

THÈSE

présentée

à la Faculté des Sciences de l'Université de Lyon

pour obtenir

le grade de Docteur du 3^e Cycle en Biologie Animale Appliquée

par

Jacques PONTIER

UNE MÉTHODE D'ANALYSE FACTORIELLE. QUELQUES APPLICATIONS A LA BIOLOGIE.

Soutenue le 2 juillet 1964

devant la commission d'examen :

MM. BRACONNIER, Président

NIGON

LEGAY

}

examineurs

J a c q u e s P O N T I E R

UNE METHODE D'ANALYSE FACTORIELLE
QUELQUES APPLICATIONS A LA BIOLOGIE

LABORATOIRE DE ZOOLOGIE EXPERIMENTALE

FACULTE DES SCIENCES DE LYON

UNIVERSITE DE LYON
FACULTE DES SCIENCES

Doyens Honoraires

MM. LONGCHAMBON, DOUIN, MOUSSA, GAUTHIER Henri.

Professeurs Honoraires

MM. LOCQUIN, de LARAMBERGUE, SOLLAUD, MENTZER, MOUSSA, JAFFARD, DOUIN,
Henri EYRAUD, DEJARDIN, VIRET.

Doyen

M. BRACONNIER

Assesseurs

MM. MESNARD
BOIDIN

Professeurs

MM. LONGCHAMBON	Minéralogie
PRETTRE	Chimie industrielle
COURTY	Chimie-Physique
KUHNER	Botanique
MALECOT	Mécanique rationnelle et appliquée
PARIS	Chimie minérale
AUMERAS	Chimie
COLONGE	Chimie organique
JANIN	Physique générale
FLANDRIN	Géologie appliquée
BERNARD	Physique industrielle
DUFAY Jean	Astronomie
NIGON	Zoologie expérimentale
GAUTHIER Henri	Géologie générale
WAUTIER	Zoologie
CHOPIN	Chimie biologique
CUEILLERON	Chimie
BRACONNIER	Calcul différentiel et intégral
EYRAUD Charles	Chimie-Physique appliquée
SARAZIN	Physique nucléaire
BOIDIN	Microbiologie et Mycologie
PERES	Physiologie

MM. MESNARD	Electronique
GRANDMONTAGNE	Physique
BOUTILLIER	Minéralogie
LAFOUCRIERE	Physique nucléaire
DREUX	Chimie organique
TEICHNER	Chimie industrielle
PAVANS DE CECCATTY	Zoologie
MERLIN	Chimie analytique
DUFAY Maurice	Physique de l'Atmosphère
FRAISSE	Mathématiques
LUBET	Physiologie
PHILBERT	Physique nucléaire
COHEN-ADAD	Chimie
DAVID Louis	Géologie générale
CHANEL	Psychophysiologie
MAURY	Méthodes Mathématiques de la Physique
DEPRAZ	Physique nucléaire
GAUTHIER Jean	Chimie

Professeurs sans chaire

MM. GAUME	Physique
FERON	Mathématiques
LAFORGUE	Physique
LEGAY	Zoologie expérimentale S.P.C.N.
UZAN	Physique S.P.C.N.
CHENEVOY	Géologie appliquée
GINET	Zoologie
ASCH	Physique
MICHEL Pierre	Minéralogie
DUCET	Biologie végétale
Mle MAITROT	Physique
MM. GEVIN	Géologie générale
PONASSE	Mathématiques
d'INCAN	Physique
LEFEBVRE	Mathématiques
PERNET	Mathématiques
CHERY	Physique
BERGEON	Physique C.S.U.
Mle GAUTHERON	Chimie biologique

Maîtres de Conférences

Mme ERICSON	Physique générale
MM. MARTEL	Mécanique
MATHIEU	Mécanique des Fluides
ERBEIA	Physique
DESCOTES	Chimie organique
PICARD Philippe	Mathématiques
LEMOIGNE	Botanique
DEMARCO	Géologie générale



MICHEL Georges
LEBRETON
DAVID Jean
TOUSSET
FEUVRAIS
PICARD Claude
SCHMITT

Biochimie
Biologie végétale
Biologie animale
Chimie propédeutique
Physique Electricité
Mathématiques
Physique M.P.C.

Maître de Conférences adjoint

M. PIERRON

Chimie

Conseiller administratif

M. BURNET

Secrétaire Principal Honoraire

M. ROUX

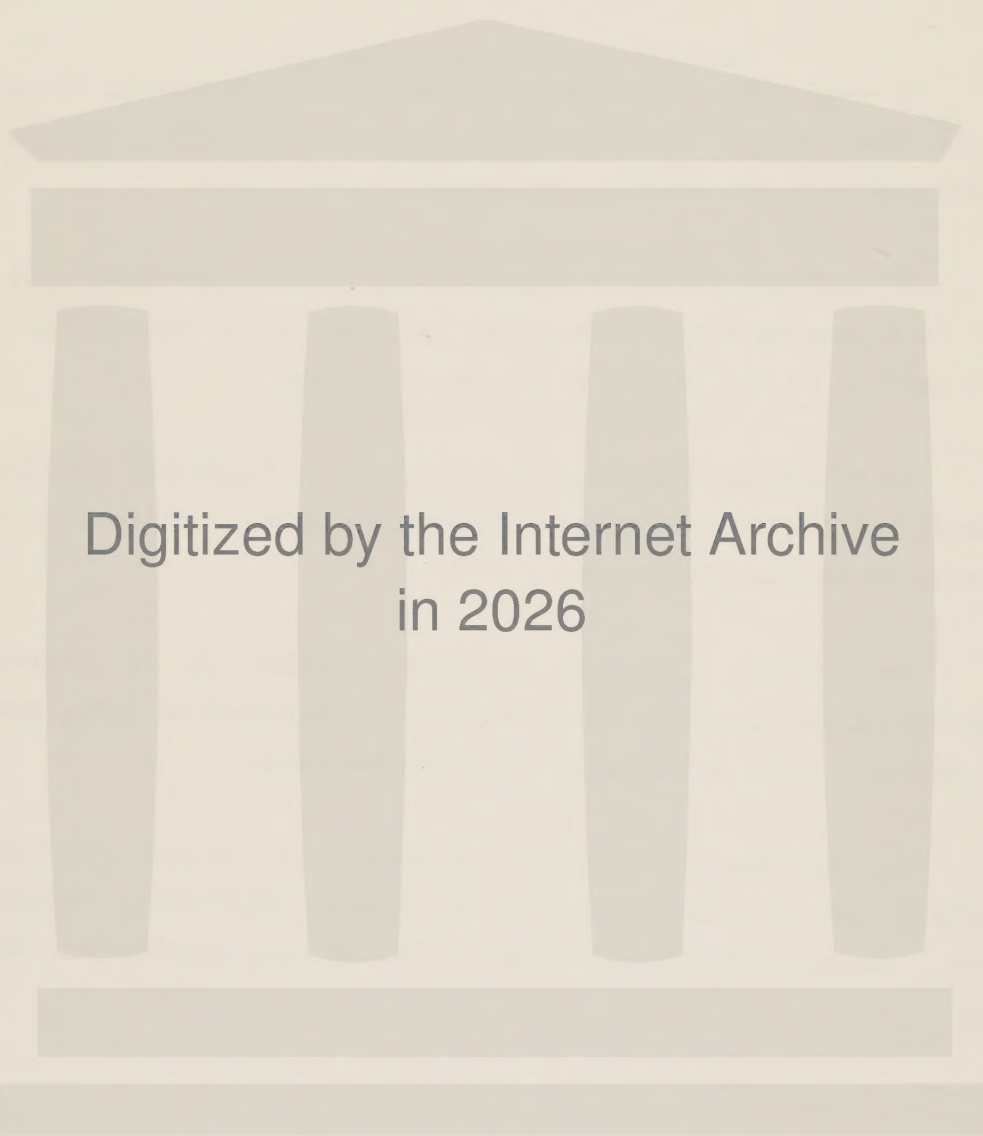
Je suis très reconnaissant à Monsieur le Professeur BRACONNIER, Doyen de la Faculté des Sciences de l' Université de Lyon, d' avoir bien voulu s'intéresser à mon travail, et accepter de le juger.

Je remercie vivement Monsieur le Professeur NIGON pour l'intérêt qu' il a accordé à mes recherches, et Monsieur le Professeur LEGAY, qui les a guidées et m'a permis de les mener à bien.

La partie mathématique de ce travail a bénéficié des conseils de Messieurs les Professeurs KAHANE, MALECOT et PICARD. Qu'ils trouvent ici l'expression de ma reconnaissance.

Je remercie également Monsieur REYNIER et le personnel du Centre de Calcul Scientifique de la Faculté des Sciences de Lyon, pour leur précieux concours lors de certains calculs numériques.

La réalisation du présent ouvrage a été rendue possible grâce à l'amabilité et à l'aide technique de Madame SZAFRANEK, Mademoiselle BRANDOLINI, Madame GUERRIER et Madame BEY. Je leur exprime ici toute ma gratitude.



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41553420_0022

S O M M A I R E

	Pages
INTRODUCTION	3
PREMIERE PARTIE : LA NOTION D'INERTIE STATISTIQUE	
CHAPITRE I : Une interprétation géométrique de l'échantillonnage	5
CHAPITRE II : Les ellipsoïdes d'inertie	17
CHAPITRE III : Quelques propriétés des axes centraux d'inertie	24
DEUXIEME PARTIE : INERTIE STATISTIQUE ET ANALYSE FACTORIELLE	
CHAPITRE I : Aperçu sur l'analyse factorielle	31
CHAPITRE II : Utilisation de l'inertie statistique	37
CHAPITRE III : Analyse factorielle et biométrie	48
TROISIEME PARTIE : EXEMPLES NUMERIQUES TIRES DE LA BIOLOGIE	
CHAPITRE I : Analyse d'un ensemble de caractères chez un Puceron	53
CHAPITRE II : Les facteurs de la forme de la feuille de Vigne	67
CHAPITRE III : Les facteurs d'hétérogénéité de l'espèce <u>Streptococcus faecalis</u>	78

CONCLUSIONS GENERALES

91

APPENDICE : QUELQUES ASPECTS TECHNIQUES

CHAPITRE I : Comparaison des éléments propres
de quelques matrices 95

CHAPITRE II : Une méthode pratique de détermi-
nation des éléments propres 101

CHAPITRE III : Récapitulation du déroulement de
l'analyse factorielle 108

BIBLIOGRAPHIE

III

I N T R O D U C T I O N

Dans le cadre assez général de l' utilisation des techniques statistiques en Biométrie, nous avons été amenés à nous intéresser aux possibilités que pouvait offrir l' Analyse Factorielle dans ce domaine. Notre point de vue était donc, à l'origine, celui de l'utilisateur, et non celui du théoricien.

Une prospection parmi les classiques de l' analyse factorielle (THURSTONE, THOMSON, etc...) nous a très vite permis de faire deux constatations assez désagréables :

- il existe de très nombreuses méthodes d'analyse factorielle, d'où embarras du choix ;

- aucune de ces méthodes (sauf peut-être celle d'HOTELLING) ne semble avoir de fondement mathématique bien précis ; en ce sens nous dirons qu'il s'agit de "recettes" plutôt que de "méthodes".

Le besoin de clarifier cette situation nous a conduits tout naturellement à nous pencher sur les fondements de l' Analyse Factorielle, afin de pouvoir utiliser cette dernière à bon escient.

Plutôt que de procéder à un examen critique systématique de chaque méthode connue - examen qui aurait été très consistant sans pour cela être nécessairement fécond - nous avons préféré forger nos propres outils, en l'occurrence notre propre méthode d'analyse factorielle.

D'aucuns diront que cela ne fait qu'une méthode de plus. En fait, si nous avons procédé ainsi, ce n'est pas pour avoir le plaisir d' ajouter une ligne à la liste déjà longue des méthodes d'analyse factorielle, mais pour disposer d'une méthode dont nous connaissons l' exacte justification mathématique (ce qui, dans une certaine mesure, nous permettra d'en connaître les possibilités et les limites d' application) ; cette méthode, de

plus, sera conçue de façon à éliminer le plus possible toute indétermination, tout choix arbitraire, pour conduire à une solution unique dans le cas général.

L'idée directrice de cette méthode est simple, mais non immédiate. Il nous a fallu, avant de l'exposer, introduire, dans la Première Partie de ce travail, ce que nous avons appelé l'"inertie statistique" (pour des raisons d'analogie profonde avec la Mécanique Rationnelle).

L'inertie est en effet une notion familière en Mécanique, où les moments et produits d'inertie, les tenseurs et les ellipsoïdes d'inertie jouent des rôles très importants. Dans cette Première Partie, nous reprenons cette notion d'inertie, en la généralisant (à n dimensions, alors que la Mécanique classique se limite à trois dimensions), et en la transposant à la Statistique. Cette transposition n'a d'ailleurs rien d'étonnant : Géométrie des masses d'une part, Calcul des Probabilités d'autre part, peuvent être considérés comme deux particularisations de la Théorie de la Mesure.

Du reste, en calcul des Probabilités, il est courant de remarquer le parallèle existant entre une distribution de masse et une distribution de probabilité, entre un centre de gravité et un point moyen (interprétation "mécaniste" des Probabilités).

En introduisant cette notion d'inertie statistique, notre but est de montrer (ce qui sera fait dans la Deuxième Partie) comment on peut utiliser, en Analyse Factorielle, les axes de ce que nous appellerons l'ellipsoïde centré d'inertie d'un échantillon. Dans cette Deuxième Partie, nous donnerons également un aperçu sur l'Analyse Factorielle en général, et sur les possibilités théoriques de l'Analyse Factorielle en Biométrie. Nous aurons aussi l'occasion d'examiner les liens de parenté existant entre notre méthode et celle d'HOTELLING (composantes principales).

Dans une Troisième Partie enfin, nous reprendrons notre point de vue initial, qui est celui du Biométricien utilisateur de l'Analyse Factorielle. Nous appliquerons notre méthode à quelques cas numériques empruntés à la Biométrie, et nous essaierons de tirer des résultats des conclusions valables du point de vue biologique.

Il nous a semblé utile d'adjoindre à cet exposé un Appendice où nous donnons quelques indications techniques, non indispensables, mais qui peuvent être de bons auxiliaires de travail pour un éventuel utilisateur de notre méthode.

P R E M I E R E P A R T I E

L A N O T I O N D ' I N E R T I E S T A T I S T I Q U E

C H A P I T R E I : U N E I N T E R P R E T A T I O N G E O M E T R I Q U E D E L ' E C H A N T I L L O N N A G E

A - V A R I A B L E S A L E A T O I R E S S I M U L T A N E E S . U N E R E P R E S E N T A T I O N G E O M E T R I Q U E D E L ' E C H A N T I L L O N

Considérons n variables aléatoires numériques simultanées $X_1, X_2, \dots, X_n$. Par exemple, dans un cas qui nous intéressera plus loin, celui de la Biométrie, ces n variables pourront être les mesures de n caractères quantitatifs distincts choisis sur les individus composant une population donnée (poids, dimensions, durées, etc...).

L'ensemble de ces n variables, soit $X = (X_1, X_2, \dots, X_n)$, peut être considéré lui-même comme une variable aléatoire à n dimensions (alors que chaque X_k est une variable aléatoire à une dimension).

Supposons que nous procédions à une étude par échantillonnage de ces n variables aléatoires. Cela signifie, par exemple, que nous extrayons "au hasard" un certain nombre N d'individus de la population considérée, et que, sur chacun de ces individus, nous mesurons les n caractères $X_1, X_2, \dots, X_n$ étudiés. Une autre façon de procéder à l'échantillonnage serait de numérotter tous les individus de la population (ou tout au moins d'une large fraction de la population), et d'opérer parmi les individus ainsi numérotés un tirage au sort, à l'aide d'une table de "nombres au hasard" (FISHER and YATES); on extrait ainsi, comme précédemment, N individus de la population, mais ici il n'est pas exclu qu'un même individu puisse figurer plusieurs fois parmi les N individus extraits.

Un tel ensemble de N individus choisis dans une population sera appelé un échantillon de taille N de la variable aléatoire X , dans la population étudiée. Sur chaque individu de l'échantillon, on mesure donc les caractères $X_1, X_2, \dots, X_n$. Les valeurs numériques de ces mesures effectivement réalisées, seront des nombres notés $x_1, x_2, \dots, x_n$. Plus précisément, pour le j -ième individu, $1 \leq j \leq N$, les valeurs de ces mesures seront notées $x_{1j}, x_{2j}, \dots, x_{nj}$. La famille $\xi_j = (x_{1j}, x_{2j}, \dots, x_{nj})$ de ces n mesures est, par extension, la "valeur" de la variable aléatoire X pour le j -ième individu de l'échantillon.

La suite de ce développement est une description des opérations que nous allons faire subir à un échantillon de taille N d'une variable aléatoire à n dimensions. La méthode d'obtention d'un tel échantillon - méthode qui pourra être l'une des deux décrites plus haut - n'a aucune influence sur ce qui va suivre.

Représentation géométrique : Nous sommes donc en présence d'un échantillon de taille N d'une certaine variable aléatoire

$$X = (X_1, X_2, \dots, X_n)$$

à n dimensions. Soit d'autre part l'espace euclidien R^n (c'est-à-dire : R^n est l'espace affine réel à n dimensions, muni de la métrique euclidienne), dans lequel nous avons choisi une origine O et un système de n axes orthonormés $Ox_1, Ox_2, \dots, Ox_n$, le tout formant un système de référence dans l'espace R^n .

A l'élément $\xi_j = (x_{1j}, x_{2j}, \dots, x_{nj})$ de l'échantillon, nous pouvons faire correspondre un point M_j de l'espace R^n : le point ayant pour coordonnées, par rapport au système d'axes choisi, les nombres $x_{1j}, x_{2j}, \dots, x_{nj}$. Ainsi, à l'échantillon de taille N , correspond un ensemble de N points de l'espace R^n ; cet ensemble de points, noté E , est habituellement appelé un "nuage statistique" ; il est l'"image" de l'échantillon par cette représentation géométrique. Dans la suite, la distinction à faire entre l'échantillon et le nuage statistique n'étant pas fondamentale, nous désignerons aussi l'échantillon lui-même par la lettre E ; il n'y aura pas d'ambiguïté. (Voir figure 1, page 30).

Nous compléterons cette représentation de l'échantillon en attribuant à chacun des points M_j ainsi définis, un certain coefficient m_j (appelé masse de M_j), qui dans le cas général sera un nombre réel positif, et

qu'ici nous ferons, par convention, égal à $\frac{1}{N}$, quel que soit $j = 1, 2, \dots$
.., N.

La masse d'un ensemble de points sera, par définition, la somme des masses de ces points. En particulier la masse du nuage statistique E tout entier est égale à l'unité.

B - INTERPRETATION DES MOYENNES D'ECHANTILLON DE $X_1, X_2, \dots, X_n$

La moyenne d'échantillon de la variable X_k , dans l'échantillon considéré, est, par définition, le nombre

$$\bar{x}_k = \frac{1}{N} \sum_{j=1}^N x_{kj}$$

$\bar{x}_k$ est une "estimation" de la moyenne vraie de X_k , soit

$$\mu_k = \mu(X_k).$$

D'autre part, en géométrie affine élémentaire, on définit le barycentre de N points M_j affectés des masses m_j : c'est le point G tel que

$$\left(\sum_{j=1}^N m_j \right) \overrightarrow{OG} = \sum_{j=1}^N m_j \overrightarrow{OM_j}$$

$$(\text{sous réserve que } \sum_{j=1}^N m_j \neq 0).$$

Le point G ainsi défini peut sembler dépendre non seulement des points M_j et des coefficients m_j , mais aussi du point O. En réalité, on complète cette définition en démontrant que le point G ainsi défini ne dépend pas du point O. Ce point G a ainsi une signification intrinsèque, il est associé d'une manière unique à l'ensemble des points M_j affectés des masses m_j .

Dans le cas particulier où $m_j = \frac{1}{N}$ quel que soit j, ce qui est le cas ici, cette définition du barycentre s'écrit :

$$\vec{OG} = \frac{1}{N} \sum_{j=1}^N \vec{OM}_j$$

Le barycentre est encore appelé, dans ce cas, "centre de gravité" ou "centre d'inertie". Dans ce qui suit, nous utiliserons l'expression "centre d'inertie".

La k-ième composante du vecteur $\vec{OG}$, c'est-à-dire la k-ième coordonnée du point G, est alors égale à

$$\frac{1}{N} \sum_{j=1}^N x_{kj}$$

ce qui n'est autre que la moyenne d'échantillon de X_k dans l'échantillon considéré E, soit $\bar{x}_k$. Donc :

Proposition 1 : Les n moyennes d'échantillon $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$ sont les coordonnées du centre d'inertie G du nuage statistique E. (Il est sous-entendu, rappelons-le, qu'il s'agit de coordonnées par rapport au système d'axes $Ox_1, Ox_2, \dots, Ox_n$ choisi au paragraphe A ci-dessus).

C - MOMENTS D'INERTIE. INTERPRETATION DES VARIANCES ET COVARIANCES

Soit V une variété linéaire dans R^n euclidien (V peut être un point, ou une droite, ..., ou un hyperplan de R^n).

Notons d_j la distance euclidienne du point M_j à la variété linéaire V :

$$d_j = \text{dist. } (M_j, V)$$

Nous appellerons moment d'inertie, par rapport à V, de l'ensemble E de points M_j affectés des masses m_j , le nombre

$$I(E/V) = \sum_{j=1}^N m_j d_j^2$$

Dans le cas qui nous intéresse ici, $m_j = \frac{1}{N}$, ce moment d'inertie s'écrit :

$$I(E/V) = \frac{1}{N} \sum_{j=1}^N d_j^2$$

Dans ce cas $I(E/V)$ peut aussi être interprété comme le carré de la distance quadratique moyenne des points de E à la variété linéaire V.

Notons au passage la propriété suivante :

Proposition 2 : $I(E/V)$ est nul si et seulement si tous les points de E appartiennent à la variété linéaire V (car alors $d_j = 0$ quel que soit j).

Exemples de moments d'inertie :

a) moment d'inertie par rapport à un point :

V est une variété linéaire de dimension zéro, c'est-à-dire un point M_0 , de coordonnées $x_{10}, x_{20}, \dots, x_{n0}$. Le carré de la distance d_j est

$$d_j^2 = (x_{1j} - x_{10})^2 + (x_{2j} - x_{20})^2 + \dots + (x_{nj} - x_{n0})^2$$

D'où :

$$I(E/M_0) = \frac{1}{N} \sum_{j=1}^N (x_{1j} - x_{10})^2 + \dots + \frac{1}{N} \sum_{j=1}^N (x_{nj} - x_{n0})^2$$

Un cas particulier intéressant est celui où le point M_0 est le centre d'inertie G du système E ; on a alors

$$I(E/G) = \frac{1}{N} \sum_{j=1}^N (x_{1j} - \bar{x}_1)^2 + \dots + \frac{1}{N} \sum_{j=1}^N (x_{nj} - \bar{x}_n)^2$$

Or, par définition, la quantité $s_k^2 = \frac{1}{N} \sum_{j=1}^N (x_{kj} - \bar{x}_k)^2$ est la variance d'échantillon de X_k dans l'échantillon E.

Proposition 3 : Le moment d'inertie de E par rapport au centre d'inertie G de E est la somme $s_1^2 + s_2^2 + \dots + s_n^2$ des variances d'échantillon de $X_1, X_2, \dots, X_n$.

Pour désigner cette somme des n variances, nous utiliserons ci-après la notation $s_E^2 = s_1^2 + s_2^2 + \dots + s_n^2$.

b) moment d'inertie par rapport à une variété linéaire à r dimensions, passant par $M_O(x_{10}, x_{20}, \dots, x_{no})$ parallèle à $Ox_1, Ox_2, \dots, Ox_r$:

Désignons par V cette variété linéaire. Un point M_j de E ayant pour coordonnées $x_{1j}, x_{2j}, \dots, x_{nj}$, sa projection orthogonale M'_j dans V a pour coordonnées :

$$x_{1j}, x_{2j}, \dots, x_{rj}, x_{r+1,0}, x_{r+2,0}, \dots, x_{no}$$

Les composantes du vecteur $\overrightarrow{M'_j M_j}$ sont donc :

$$0, 0, \dots, 0, x_{r+1,j} - x_{r+1,0}, \dots, x_{nj} - x_{no}$$

D'où :

$$d_j^2 = \overrightarrow{M'_j M_j}^2 = (x_{r+1,j} - x_{r+1,0})^2 + \dots + (x_{nj} - x_{no})^2, \text{ et}$$

$$I(E/V) = \frac{1}{N} \sum_{j=1}^N (x_{r+1,j} - x_{r+1,0})^2 + \dots + \frac{1}{N} \sum_{j=1}^N (x_{nj} - x_{no})^2$$

En particulier, si le point M est le centre d'inertie G de E , on a :

$$I(E/V) = \frac{1}{N} \sum_{j=1}^N (x_{r+1,j} - \bar{x}_{r+1})^2 + \dots + \frac{1}{N} \sum_{j=1}^N (x_{nj} - \bar{x}_n)^2$$

c'est-à-dire :

Proposition 4 : V étant la variété linéaire passant par G , parallèle à $Ox_1, Ox_2, \dots, Ox_r$, le moment d'inertie de E par rapport à V est :

$$I(E/V) = s_{r+1}^2 + s_{r+2}^2 + \dots + s_n^2 = s_E^2 - (s_1^2 + s_2^2 + \dots + s_r^2)$$

De cette proposition découle immédiatement le corollaire suivant :

Proposition 5 : H_k étant l'hyperplan passant par G , orthogonal à Ox_k ,
on a :

$$I(E/H_k) = s_k^2$$

La variance d'échantillon de X_k peut donc être interprétée comme un moment d'inertie.

c) moment d'inertie par rapport à une droite :

V est maintenant une variété linéaire à une dimension, c'est-à-dire une droite D . Elle est définie par l'un de ses points, soit M_0 , de coordonnées $x_{10}, x_{20}, \dots, x_{n0}$, et par ses n cosinus directeurs $\alpha_1, \alpha_2, \dots, \alpha_n$ (qui sont les n composantes d'un vecteur unitaire $\vec{V}$ parallèle à D).

Soit M'_j la projection orthogonale de M_j sur D . Nous avons (voir figure 2)

$$|M_0 M'_j| = |\overrightarrow{M_0 M_j} \cdot \vec{V}|, \text{ car } \vec{V} \text{ est unitaire. De plus :}$$

$$d_j^2 = \overrightarrow{M_j M_j}^2 = \overrightarrow{M_0 M_j}^2 - \overrightarrow{M_0 M'_j}^2$$

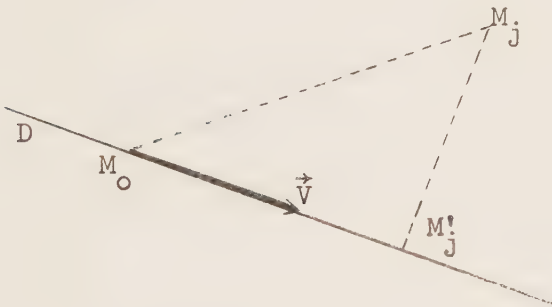


Figure 2

Or : $\overrightarrow{M_0 M_j}^2 = (x_{1j} - x_{10})^2 + \dots + (x_{nj} - x_{n0})^2$; et d'autre part :

$$\overrightarrow{M_0 M_j} \cdot \vec{V} = \alpha_1 (x_{1j} - x_{10}) + \alpha_2 (x_{2j} - x_{20}) + \dots + \alpha_n (x_{nj} - x_{n0}) ; \text{ d'où}$$

$$\overrightarrow{M_0 M'_j}^2 = \alpha_1^2 (x_{1j} - x_{10})^2 + \dots + \alpha_n^2 (x_{nj} - x_{n0})^2 + 2\alpha_1 \alpha_2 (x_{1j} - x_{10})$$

$$(x_{2j} - x_{20}) + \dots + 2\alpha_{n-1} \alpha_n (x_{n-1,j} - x_{n-1,0})(x_{nj} - x_{n0})$$

Nous pouvons alors exprimer $d_j^2 = \overrightarrow{M_0 M_j}^2 - \overrightarrow{M_0 M'_j}^2$. Remarquons cependant que l'on a la relation $\alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2 = 1$ entre les cosinus direc-

teurs de D. $\overrightarrow{OM_j}^2$ peut alors s'écrire sous la forme :

$$\overrightarrow{OM_j}^2 = (\alpha_1^2 + \dots + \alpha_n^2) ((x_{1j} - x_{10})^2 + \dots + (x_{nj} - x_{n0})^2)$$

D'où :

$$d_j^2 = \alpha_1^2 ((x_{2j} - x_{20})^2 + \dots + (x_{nj} - x_{n0})^2) + \dots + \alpha_n^2 ((x_{1j} - x_{10})^2 + \dots + (x_{n-1,j} - x_{n-1,0})^2) - 2\alpha_1\alpha_2 (x_{1j} - x_{10})(x_{2j} - x_{20}) - \dots$$

A partir de là on peut facilement calculer le moment d'inertie $I(E/D)$. Le cas qui nous intéressera est celui où le point M est le centre d'inertie G de E. Dans ces conditions, l'expression de $I(E/D)$ est :

$$\alpha_1^2 (s_2^2 + s_3^2 + \dots + s_n^2) + \dots + \alpha_n^2 (s_1^2 + \dots + s_{n-1}^2) - 2\alpha_1\alpha_2 \text{ cov}(X_1, X_2) - \dots - 2\alpha_{n-1}\alpha_n \text{ cov}(X_{n-1}, X_n)$$

en rappelant que, par définition, l'expression

$$\frac{1}{N} \sum_{j=1}^N (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$$

est la covariance d'échantillon de X_i et X_k dans l'échantillon E, covariance notée $\text{cov}(X_i, X_k)$ ou encore s_{ik} .

Proposition 6 : D étant une droite passant par G, de cosinus directeurs $\alpha_1, \alpha_2, \dots, \alpha_n$, on a :

$$I(E/D) = \alpha_1^2 (s_E^2 - s_1^2) + \dots + \alpha_n^2 (s_E^2 - s_n^2)$$

$$- 2\alpha_1\alpha_2 \text{ cov}(X_1, X_2) - \dots - 2\alpha_{n-1}\alpha_n \text{ cov}(X_{n-1}, X_n) .$$

Cas particulier : si D, passant par G, est parallèle à l'un des axes du système de référence, soit Ox_k , on a :

$$\left\{ \begin{array}{l} \alpha_1 = \alpha_2 = \dots = \alpha_{k-1} = \alpha_{k+1} = \dots = \alpha_n = 0 \\ \alpha_k = 1 \end{array} \right.$$

$$D'où : I(E/D) = s_E^2 - s_k^2 = s_1^2 + \dots + s_{k-1}^2 + s_{k+1}^2 + \dots + s_n^2$$

Ce résultat était d'ailleurs prévisible comme conséquence de la proposition 4.

D - TENSEUR D'INERTIE EN UN POINT

Dans l'expression générale du moment d'inertie $I(E/D)$, où D est une droite quelconque passant par M_0 , nous conviendrons de noter :

$$I_{pp} = \frac{1}{N} \sum_{j=1}^N ((x_{1j} - x_{10})^2 + \dots + (x_{p-1,j} - x_{p-1,0})^2 + (x_{p+1,j} - x_{p+1,0})^2 + \dots + (x_{nj} - x_{no})^2)$$

I_{pp} est le moment d'inertie de E par rapport à la droite D_p , passant par M , parallèle à Ox_p (voir démonstration de la proposition 4).

Nous noterons aussi :

$$I_{pq} = \frac{1}{N} \sum_{j=1}^N (x_{pj} - x_{po}) (x_{qj} - x_{qo}) = I_{qp} ;$$

Ce nombre I_{pq} (ou I_{qp}) est appelé le produit d'inertie de E par rapport aux deux droites D_p, D_q (passant par M_0 , parallèles respectivement à Ox_p et Ox_q).

Avec ces notations, nous pouvons écrire :

$$I(E/D) = \alpha_1^2 I_{11} + \dots + \alpha_n^2 I_{nn} - 2\alpha_1 \alpha_2 I_{12} - \dots - 2\alpha_{n-1} \alpha_n I_{n-1,n}.$$

Le second membre est une forme quadratique, qui s'écrit encore sous forme matricielle :

$$I(E/D) = (\alpha_1 \alpha_2 \dots \alpha_n) \begin{pmatrix} I_{11} & -I_{12} & \dots & -I_{1n} \\ -I_{21} & I_{22} & \dots & -I_{2n} \\ \vdots & \vdots & & \vdots \\ -I_{n1} & -I_{n2} & \dots & -I_{nn} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

ou encore : $I(E/D) = {}^t_v A v$, où :

- v est la matrice associée au vecteur $\vec{v}$, dans le système de référence $Ox_1, Ox_2, \dots, Ox_n$, soit :

$$v = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

- t_v est la transposée de la matrice v , soit

$${}^t_v = (\alpha_1 \alpha_2 \dots \alpha_n)$$

- A est la matrice carrée $n \times n$, formée des moments et produits d'inertie ; c'est une matrice symétrique car $I_{pq} = I_{qp}$. Cette matrice, servant à exprimer le moment d'inertie de E par rapport à la droite D , est appelée le tenseur d'inertie de E au point M_0 (en réalité, le tenseur est un "opérateur linéaire" sur R^n , et a par conséquent une signification intrinsèque ; alors que la matrice A est la représentation matricielle de cet opérateur, par rapport au système d'axes $Ox_1, Ox_2, \dots, Ox_n$; c'est par abus de langage que nous identifions la matrice A et l'opérateur - tenseur- qu'elle représente).

E - THEOREME DE HUYGHENS

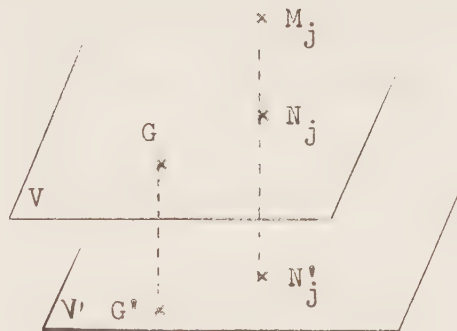
Considérons deux variétés linéaires V et V' , parallèles et de même dimension ; supposons que V passe par G , centre d'inertie de E (voir figure 3).

Peut-on exprimer l'un par rapport à l'autre les deux moments

d'inertie $I(E/V)$ et $I(E/V')$?

Soit M_j un point de E . Sa projection orthogonale sur V est notée N_j ; sa projection orthogonale sur V' est N'_j . De plus, la projection orthogonale de G sur V' est notée G' .

Figure 3



Les carrés des distances de M_j à V et V' respectivement sont :

$$d_j^2 = \overrightarrow{N_j M_j}^2 ; d'_j{}^2 = \overrightarrow{N'_j M_j}^2. \text{ Mais : } \overrightarrow{N_j M_j} = \overrightarrow{G M_j} - \overrightarrow{G N_j} ; \text{ et } \overrightarrow{N'_j M_j} = \overrightarrow{G M_j} - \overrightarrow{G N'_j} .$$

$$\text{D'où : } d_j^2 = \overrightarrow{G M_j}^2 + \overrightarrow{G N_j}^2 - 2 \overrightarrow{G M_j} \cdot \overrightarrow{G N_j} , \text{ et } d'_j{}^2 = \overrightarrow{G M_j}^2 + \overrightarrow{G N'_j}^2 - 2 \overrightarrow{G M_j} \cdot \overrightarrow{G N'_j}$$

En remarquant que : $\overrightarrow{G' M_j} = \overrightarrow{G' G} + \overrightarrow{G M_j}$, et $\overrightarrow{G' N'_j} = \overrightarrow{G N'_j}$, nous avons :

$$\begin{aligned} d'_j{}^2 &= \overrightarrow{G' G}^2 + \overrightarrow{G M_j}^2 + 2 \overrightarrow{G' G} \cdot \overrightarrow{G M_j} + \overrightarrow{G N'_j}^2 - 2 \overrightarrow{G' G} \cdot \overrightarrow{G N'_j} - 2 \overrightarrow{G M_j} \cdot \overrightarrow{G N'_j} \\ &= d_j^2 + \overrightarrow{G' G}^2 + 2 \overrightarrow{G' G} \cdot \overrightarrow{G M_j} - 2 \overrightarrow{G' G} \cdot \overrightarrow{G N'_j} . \end{aligned}$$

Or $\overrightarrow{G' G}$ est orthogonal à V , donc à $\overrightarrow{G N_j}$; d'où : $\overrightarrow{G' G} \cdot \overrightarrow{G N_j} = 0$

Il en résulte que :

$$\begin{aligned} I(E/V) - I(E/V') &= \frac{1}{N} \sum_{j=1}^N (d_j^2 - d'_j{}^2) \\ &= \frac{1}{N} \sum_{j=1}^N (-\overrightarrow{G' G}^2 - 2 \overrightarrow{G' G} \cdot \overrightarrow{G M_j}) \\ &= -\overrightarrow{G' G}^2 - \frac{2}{N} \overrightarrow{G' G} \cdot \sum_{j=1}^N \overrightarrow{G M_j} \end{aligned}$$

Mais, d'après la définition du centre d'inertie, nous avons :

$$\frac{1}{N} \sum_{j=1}^N \overrightarrow{G M_j} = \vec{O} . \text{ D'où : } I(E/V') = I(E/V) + \overrightarrow{G' G}^2$$

Remarquons que $\overrightarrow{G'G}^2$ est le carré de la distance des deux variétés linéaires parallèles V et V' .

Proposition 7 (Théorème de HUYGHENS) : Si V et V' sont deux variétés linéaires parallèles et de même dimension, et si V passe par le centre d'inertie G de E , on a la relation :

$$I(E/V') = I(E/V) + a^2, \text{ avec } a = \text{dist.}(V, V')$$

Ce résultat est d'ailleurs utilisé dans le calcul pratique d'une variance. En effet, nous avons vu plus haut (proposition 5) qu'une variance peut être interprétée comme un moment d'inertie par rapport à un hyperplan :

$$s_k^2 = I(E/H_k), \quad H_k \text{ hyperplan passant par } G, \text{ orthogonal à } Ox_k.$$

Soit H'_k l'hyperplan passant par O , orthogonal à Ox_k (donc parallèle à H_k).

$$\text{Nous avons : } I(E/H'_k) = \frac{1}{N} \sum_{j=1}^N (x_{kj})^2 = S_k^2$$

et d'autre part $\text{dist.}(H_k, H'_k) = |\bar{x}_k|$; d'où :

$$I(E/H'_k) = I(E/H_k) + \bar{x}_k^2, \quad \text{c'est-à-dire :}$$

$$s_k^2 = \frac{1}{N} \sum_{j=1}^N x_{kj}^2 - \bar{x}_k^2,$$

formule bien connue utilisée pratiquement pour le calcul de s_k^2 .

C H A P I T R E I I : L E S E L L I P S O I D E S D ' I N E R T I E

A - ELLIPSOIDE D'INERTIE EN UN POINT

Considérons, dans l'espace euclidien R^n où est représenté l'échantillon E , un point quelconque $M_O (x_{10}, x_{20}, \dots, x_{n0})$, un vecteur unitaire $\vec{v}$ de composantes $\alpha_1, \alpha_2, \dots, \alpha_n$, et une droite D , passant par M_O , parallèle au vecteur $\vec{v}$.

Nous avons vu (voir démonstration de la proposition 6) que le moment d'inertie de E , nuage statistique, par rapport à la droite D , s'exprime sous la forme :

$$I(E/D) = \alpha_1^2 I_{11} + \dots + \alpha_n^2 I_{nn} - 2\alpha_1\alpha_2 I_{12} - \dots - 2\alpha_{n-1}\alpha_n I_{n-1,n}$$

Une variation de la direction du vecteur $\vec{v}$ entraîne une variation de la direction de la droite D , et par conséquent aussi une éventuelle variation de la quantité $I(E/D)$.

Nous nous proposons d'étudier les variations de $I(E/D)$ lorsque le vecteur $\vec{v}$ (ou la droite D) varie, le point M_O restant fixé au cours de ces variations. Ce moment d'inertie $I(E/D)$ apparaît comme une fonction réelle des n variables $\alpha_1, \alpha_2, \dots, \alpha_n$ (en fait, il n'y a que $n-1$ variables indépendantes puisque $\alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2 = 1$). Mais l'étude de $I(E/D)$ comme fonction de n variables n'est pas ici d'un grand intérêt. Nous préférons procéder à une étude "géométrique" de la variation de $I(E/D)$, en construisant dans l'espace une indicatrice du moment d'inertie de E par rapport à D ; ceci, de la manière suivante :

Remarquons d'abord que $I(E/D)$ est une quantité essentiellement positive (somme de carrés de nombres réels), pouvant exceptionnellement être nulle (voir proposition 2).

Dans tous les cas $\sqrt{I(E/D)}$ existe.

Sur la droite D , considérons le point P défini par la relation vectorielle :

$$\overrightarrow{M_O P} = \frac{1}{\sqrt{I(E/D)}} \vec{v}$$

C'est dire simplement que la longueur du vecteur $\overrightarrow{M_0 P}$ est égale à

$$\frac{1}{\sqrt{I(E/D)}}$$

Lorsque le vecteur $\vec{v}$ varie, le point P se déplace dans l'espace R^n ; l'ensemble des positions prises par le point P (lieu géométrique) constitue l'indicatrice que nous recherchons, et que nous allons étudier de plus près.

Remarquons tout d'abord que M_0 est centre de symétrie de cette indicatrice. En effet, aux deux vecteurs opposés $\vec{v}$ et $-\vec{v}$ correspond une même droite D, et par conséquent une même valeur de $I(E/D)$; les deux points P et P' correspondants sont donc symétriques par rapport à M_0 : $\overrightarrow{M_0 P} = - \overrightarrow{M_0 P'}$.

Recherchons maintenant l'équation de l'indicatrice. Désignons par $x_1, x_2, \dots, x_n$ les coordonnées du point P. La relation vectorielle définissant P se traduit alors par les n relations :

$$x_j - x_{j0} = \frac{\alpha_j}{\sqrt{I(E/D)}} \quad j = 1, 2, \dots, n$$

d'où nous déduisons :

$$\left\{ \begin{array}{l} \alpha_j^2 = I(E/D) (x_j - x_{j0})^2 \quad j = 1, 2, \dots, n \\ \alpha_j \alpha_k = I(E/D) (x_j - x_{j0}) (x_k - x_{k0}) \quad j, k = 1, 2, \dots, n \end{array} \right.$$

L'expression de $I(E/D)$ devient alors :

$$I(E/D) = I(E/D) \left\{ I_{11} (x_1 - x_{10})^2 + \dots + I_{nn} (x_n - x_{n0})^2 - 2 I_{12} (x_1 - x_{10}) (x_2 - x_{20}) - \dots \right\}$$

Dans le cas que nous rencontrerons le plus généralement, nous aurons $I(E/D) \neq 0$ quelle que soit la droite D ; ceci autorise la division par $I(E/D)$ des deux membres de la relation ci-dessus, qui devient alors :

$$I_{11} (x_1 - x_{10})^2 + \dots + I_{nn} (x_n - x_{n0})^2 - 2 I_{12} (x_1 - x_{10}) (x_2 - x_{20}) - \dots = 1$$

Cette dernière relation, liant les coordonnées $x_1, x_2, \dots, x_n$ de P, est l'équation de l'indicatrice. Nous reconnaissons l'équation d'une quadrique à n dimensions. Cette quadrique a la particularité suivante : la distance de l'un quelconque de ses points, soit P, au point M_0 , est finie (égale par construction à $\frac{1}{\sqrt{I(E/D)}}$, où D est la droite M_0P et où $I(E/D) \neq 0$).

Cette quadrique est donc un ellipsoïde, que nous appellerons l'ellipsoïde d'inertie de E en M_0 .

Cas particuliers : ce sont les cas où il existe une (ou des) droite (s), telle(s) que $I(E/D) = 0$.

- a) s'il existe une seule droite D_0 telle que $I(E/D_0) = 0$, cela signifie que tous les points de E sont sur la droite D_0 . L'ellipsoïde d'inertie est dans ce cas une surface cylindrique d'axe D_0 .
- b) s'il existe deux droites D_0 et D_1 , telles que $I(E/D_0) = I(E/D_1) = 0$, cela entraîne que tous les points de E sont situés à la fois sur D_0 et sur D_1 ; c'est dire que E est réduit à un point, qui est l'intersection M_0 de ces deux droites (notons qu'alors M_0 n'est autre que le point G, centre d'inertie). Dans ces conditions, on a $I(E/D) = 0$ quelle que soit la droite D passant par M_0 . On peut dire alors que l'ellipsoïde d'inertie est une hypersphère centrée en M_0 , de rayon infini.

B - AXES PRINCIPAUX D'INERTIE EN M_0

Un ellipsoïde dans l'espace R^n admet n axes de symétrie, deux à deux orthogonaux, passant tous par le centre M_0 de cet ellipsoïde.

Les n axes de symétrie de l'ellipsoïde d'inertie en M_0 sont appelés les axes principaux d'inertie de E en M_0 .

Généralement, ces n axes sont bien déterminés : leur existence est assurée d'une manière unique ; mais il peut arriver qu'il y ait une indétermination portant sur k de ces axes ; ces derniers pourront alors être choisis assez arbitrairement, dans des conditions que nous préciserons (paragraphe D du présent chapitre).

Ainsi, à deux dimensions, l'ellipsoïde est le plus souvent une ellipse dans le plan R^2 ; ses deux axes sont bien déterminés ; mais il se

peut que cette ellipse soit un cercle, et dans ce cas on choisit les deux axes d'une manière arbitraire, pourvu qu'ils soient orthogonaux entre eux.

C - L'ELLIPSOÏDE CENTRE D'INERTIE D'UN ECHANTILLON

Nous avons défini ci-dessus l'ellipsoïde d'inertie d'un échantillon en un point M_0 . Dans le cas particulier où ce point M_0 est le centre d'inertie de E, soit G, l'ellipsoïde d'inertie correspondant est appelé l'ellipsoïde centré d'inertie de E.

L'équation de l'ellipsoïde centré d'inertie est donc, par rapport au système d'axes choisi $Ox_1, Ox_2, \dots, Ox_n$:

$$(s_E^2 - s_1^2) (x_1 - \bar{x}_1)^2 + \dots + (s_E^2 - s_n^2) (x_n - \bar{x}_n)^2$$

$$- 2 s_{12} (x_1 - \bar{x}_1) (x_2 - \bar{x}_2) - \dots = 1 ,$$

ou encore, en écriture matricielle :

$$(x_1 - \bar{x}_1, \dots, x_n - \bar{x}_n) \begin{pmatrix} s_E^2 - s_1^2 & -s_{12} & \dots & -s_{1n} \\ -s_{21} & s_E^2 - s_2^2 & \dots & -s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -s_{n1} & -s_{n2} & \dots & s_E^2 - s_n^2 \end{pmatrix} \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_n - \bar{x}_n \end{pmatrix} = 1$$

La matrice carrée symétrique $n \times n$ intervenant dans cette équation est le tenseur d'inertie de E au point G ; on l'appelle le tenseur centré d'inertie de E, noté T_G .

Remarquons que T_G s'écrit :

$$T_G = s_E^2 I - S ,$$

où I est la matrice unité d'ordre n, et S la matrice des covariances de l'échantillon, soit :

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1n} \\ s_{21} & s_2^2 & & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & & s_n^2 \end{pmatrix} \quad \begin{aligned} s_j^2 &= s_{jj} = \text{var } (x_j) \\ s_{jk} &= s_{kj} = \text{cov } (x_j, x_k) \end{aligned}$$

Les n axes principaux de l'ellipsoïde d'inertie de E sont appelés les axes centraux d'inertie de E .

D - DETERMINATION DES AXES CENTRAUX D'INERTIE D'UN ECHANTILLON

Les axes centraux d'inertie de l'échantillon E vont jouer, dans la suite de cet exposé, un rôle fondamental. Leur détermination pratique apparaîtra, pour les applications numériques de la présente théorie, comme un point très important. Or, bien que techniquement possible, cette détermination nécessite des calculs très longs dès que le nombre des variables considérées est supérieur à 4 ou 5. Heureusement, la technique vient au secours du chercheur, en mettant à sa disposition diverses machines à calculer, notamment les ordinateurs électroniques, lesquels s'avèreront, pour résoudre ce problème-là, d'un secours extrêmement précieux, sinon indispensable.

La détermination des axes centraux d'inertie est un aspect de la "réduction des formes quadratiques", question dont la résolution théorique est parfaitement au point. Il nous suffira ici de savoir que les directions de ces axes sont les "directions propres" de la matrice T_G définissant l'ellipsoïde d'inertie (PISOT et ZAMANSKY, 1961, p. 189-202).

Rappelons quelques points essentiels de cette théorie :

a) quelques définitions : $\vec{v}$ est un vecteur propre de la matrice T_G , et λ (nombre réel) est la valeur propre associée, si l'on a la relation :

$$T_G \vec{v} = \lambda \vec{v}$$

En d'autres termes, $\vec{v}$ est un vecteur propre de T_G si son transformé par la transformation T_G (tenseur d'inertie) est un vecteur de même direction que $\vec{v}$.

De plus, si $\vec{v}$ est un vecteur propre de T_G , associé à une valeur propre λ , tout vecteur $k\vec{v}$ de même direction que $\vec{v}$ est lui-même un vecteur propre de T_G , associé à la même valeur propre λ . Ainsi, tous les vecteurs de même direction que $\vec{v}$ sont des vecteurs propres associés à λ . La direction commune à ces vecteurs est appelée la direction propre associée à la valeur propre λ .

- b) quelques propriétés : la matrice T_G étant ici symétrique (donc hermitique), elle admet n valeurs propres réelles. Deux directions propres associées à deux valeurs propres distinctes sont orthogonales.

La matrice T_G possède la propriété suivante : $v'T_G v \geq 0$ quel que soit le vecteur $\vec{v}$ (on dit alors que T_G est semi-définie-positive). Cela entraîne que les n valeurs propres de T_G sont toutes positives ou nulles. En fait, si le déterminant de T_G n'est pas nul, les n valeurs propres sont strictement positives.

Envisageons ici le cas spécial dont nous avons parlé au paragraphe B du présent chapitre :

- si k valeurs propres sont égales, il correspond à leur valeur commune λ un sous-espace propre (c'est-à-dire un sous-espace dont tout vecteur est vecteur propre de T_G) associé à la valeur propre multiple λ ; ce sous-espace est de dimension k ; il est orthogonal aux directions propres associées aux $n-k$ autres valeurs propres. On choisit alors, dans ce sous-espace propre, k vecteurs orthogonaux entre eux, qui définissent k directions propres associées à la valeur propre multiple λ (ce choix est possible d'une infinité de manières).

- c) calcul numérique : on peut montrer que les valeurs propres de A , matrice carrée $n \times n$ quelconque, sont les racines de l'équation :

$$\det | A - \lambda I | = 0 ,$$

où λ est l'indéterminée, I est la matrice unité $n \times n$.

Théoriquement donc, la détermination des valeurs propres d'une matrice peut se concevoir comme la résolution d'une équation de degré n . En fait, cette méthode ne s'avère facilement utilisable que dans les cas $n = 2$ ou $n = 3$ (sauf cas particulier, bien entendu).

On préfère généralement utiliser des méthodes indirectes :
- celles où on opère sur A des transformations unitaires successives, du

type $T^{-1}AT$, jusqu'à obtenir une matrice diagonale A' , équivalente à A , dont les éléments diagonaux sont les valeurs propres cherchées (exemple: méthode de JACOBI). La connaissance des matrices unitaires T successives permet de déterminer aisément les vecteurs propres.

- celles où on procède par itération (exemple : méthode d'HOTELLING) ; nous n'insisterons pas pour le moment sur ce genre de méthode, que nous exposerons en détail plus loin (Appendice).

- celles enfin qui sont des combinaisons des deux précédentes, combinaisons destinées à minimiser le nombre d'opérations nécessaires pour arriver au résultat, et à augmenter la précision de ce dernier.

Toutes ces méthodes sont expliquées, et appliquées numériquement, dans l'ouvrage de DURAND (1962), tome 2, qui leur est spécialement consacré.

E - LA PRESENTATION DES RESULTATS (NOTATIONS)

Supposons que le problème de la réduction de l'ellipsoïde centré d'inertie ait été complètement résolu numériquement, par l'une des méthodes dont nous avons parlé ci-dessus. Nous nous trouvons alors en présence :

- 1° - d'un ensemble de n nombres réels positifs λ_j ($j = 1, 2, \dots, n$), qui sont les valeurs propres du tenseur d'inertie T_G .
- 2° - d'un ensemble de n vecteurs $\vec{v}_j$ de l'espace R^n (chaque vecteur étant déterminé par ses n composantes par rapport au système de référence initialement choisi), qui sont des vecteurs propres de T_G ($\vec{v}_j$ associé à la valeur propre λ_j).

Il sera commode, pour la suite de l'exposé :

- d'une part, d'ordonner les valeurs propres λ_j suivant les grandeurs croissantes : $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$

- d'autre part, de rendre chacun des vecteurs $\vec{v}_j$ unitaire (en divisant chacune des composantes par la longueur de ce vecteur). Les composantes du vecteur propre unitaire $\vec{v}_j$ seront désignées par $a_{j1}, a_{j2}, \dots, a_{jn}$. On notera C la matrice carrée :

$$C = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & & a_{nn} \end{pmatrix}$$

CHAPITRE III : QUELQUES PROPRIETES DES AXES CENTRAUX D'INERTIE

A - UN CHANGEMENT DE COORDONNEES DANS R^n

Les n vecteurs propres $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ de T_G formant une base dans l'espace vectoriel R^n , il nous est possible d'utiliser les n axes centraux d'inertie $Gz_1, Gz_2, \dots, Gz_n$, comme système de référence dans R^n . Nous pouvons même choisir comme vecteur unitaire, sur chaque axe Gz_k , le vecteur $\vec{v}_k$. Dans ces conditions, si $z_1, z_2, \dots, z_n$ sont les coordonnées d'un certain point M , par rapport au système $Gz_1, \dots, Gz_n$, nous pouvons écrire :

$$\vec{GM} = z_1 \vec{v}_1 + z_2 \vec{v}_2 + \dots + z_n \vec{v}_n$$

Rappelons que nous avons noté $a_{k1}, a_{k2}, \dots, a_{kn}$ les composantes du vecteur $\vec{v}_k$ par rapport au système d'axes $Ox_1, Ox_2, \dots, Ox_n$. Cela signifie que :

$$\vec{v}_k = a_{k1} \vec{u}_1 + a_{k2} \vec{u}_2 + \dots + a_{kn} \vec{u}_n$$

où $\vec{u}_j$ est le vecteur unitaire porté par Ox_j .

Donc, si $x_1, x_2, \dots, x_n$ sont les coordonnées du point M par rapport à $Ox_1, \dots, Ox_n$, nous pouvons écrire :

d'une part : $\vec{GM} = (x_1 - \bar{x}_1) \vec{u}_1 + \dots + (x_n - \bar{x}_n) \vec{u}_n$
et d'autre part :

$$\begin{aligned} \vec{GM} = & z_1 (a_{11} \vec{u}_1 + \dots + a_{1n} \vec{u}_n) + z_2 (a_{21} \vec{u}_1 + \dots + a_{2n} \vec{u}_n) + \dots \\ & \dots + z_n (a_{n1} \vec{u}_1 + \dots + a_{nn} \vec{u}_n) \end{aligned}$$

D'où la relation :

$$x_k - \bar{x}_k = z_1 a_{1k} + z_2 a_{2k} + \dots + z_n a_{nk} \quad k = 1, 2, \dots, n$$

relation qui, matriciellement, s'exprime sous la forme :

$$\begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_n - \bar{x}_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & & a_{nn} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

ou plus simplement : $X = CZ$, avec

$$X = \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_n - \bar{x}_n \end{pmatrix} ; \quad Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

(la matrice C est la matrice des a_{ij} ; cette notation a été introduite plus haut).

La relation $X = CZ$ permet le calcul des x_k , connaissant les z_j . Réciproquement, pour calculer les z_j , connaissant les x_k , il suffit d'écrire que : $Z = C^{-1}X$, C^{-1} étant la matrice inverse de la matrice C . La matrice C est connue, mais il est inutile de déterminer par le calcul la matrice C^{-1} . En effet, C est formée des vecteurs $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ unitaires et deux à deux orthogonaux. La matrice C est donc une matrice unitaire ; son inverse coïncide donc avec sa transposée : $C^{-1} = {}^t C$.

Proposition 8 : L'inverse de la matrice C est la matrice

$$C^{-1} = {}^t C = \begin{pmatrix} a_{11} & a_{12} \dots & a_{1n} \\ a_{21} & a_{22} \dots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} \dots & a_{nn} \end{pmatrix}$$

Nous pouvons résumer ce qui a été dit ci-dessus en énonçant :

Proposition 9 : Le passage du système d'axes Gz_j au système d'axes Ox_k s'effectue au moyen de la relation matricielle

$$X = CZ$$

Réciproquement, le passage du système Ox_k au système Gz_j s'effectue au moyen de la relation

$$Z = {}^tCX$$

B - EQUATION DE L'ELLIPSOÏDE CENTRE D'INERTIE, PAR RAPPORT A SES AXES

Rappelons que l'équation de l'ellipsoïde centré d'inertie de E s'écrit, par rapport aux axes $Ox_1, Ox_2, \dots, Ox_n$:

$${}^tX T_G X = 1 \quad \text{où :}$$

$$X = \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \vdots \\ x_n - \bar{x}_n \end{bmatrix} ; \quad T_G = \begin{bmatrix} s_E^2 - s_1^2 & -s_{12} & \dots & -s_{1n} \\ -s_{21} & s_E^2 - s_2^2 & \dots & -s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -s_{n1} & -s_{n2} & \dots & s_E^2 - s_n^2 \end{bmatrix}$$

Nous nous proposons de déterminer l'équation de cet ellipsoïde, par rapport aux axes $Gz_1, Gz_2, \dots, Gz_n$ définis par les vecteurs unitaires $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ déterminés plus haut.

Nous savons qu'il existe, entre les "anciennes" coordonnées $x_1, x_2, \dots, x_n$ d'un point quelconque de l'espace, et ses "nouvelles" coordonnées $z_1, z_2, \dots, z_n$, la relation

$$X = CZ \quad \text{où}$$

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} ; \quad C = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

L'équation cherchée s'écrit donc, en remplaçant X par CZ et tX

par ${}^t(CZ) = {}^tZ {}^tC$ dans l'équation initiale

$$({}^tZ {}^tC)T_G(CZ) = 1$$

ou encore ${}^tZ({}^tC {}^tT_G C)Z = 1$

où la matrice $\theta_G = {}^tC {}^tT_G C$ représente le tenseur centré d'inertie de E, exprimé cette fois par rapport au système de référence $Gz_1, Gz_2, \dots, Gz_n$.

Voyons de plus près quelle est la forme de cette matrice θ_G .

$$\theta_G = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix} \begin{pmatrix} s_E^2 - s_1^2 & -s_{12} & \dots & -s_{1n} \\ -s_{21} & \dots & \dots & \vdots \\ \vdots & & & \vdots \\ -s_{n1} & \dots & \dots & s_E^2 - s_n^2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & & a_{nn} \end{pmatrix}$$

Le développement de ce produit de trois matrices nous permet de trouver une matrice carrée $n \times n$, s'écrivant sous la forme :

$$\theta_G = \begin{pmatrix} {}^t v_i T_G v_j \end{pmatrix}$$

Rappelons que :

$$\left\{ \begin{array}{l} T_G v_j = \lambda_j v_j \quad (\vec{v}_j \text{ vecteur propre de } T_G, \lambda_j \text{ valeur propre}) \\ {}^t v_j v_j = 1 \quad (\vec{v}_j \text{ vecteur unitaire}) \\ {}^t v_i v_j = 0 \text{ si } i \neq j \quad (\vec{v}_i, \vec{v}_j \text{ orthogonaux}) \end{array} \right.$$

$$\text{Donc : } {}^t v_i T_G v_j = \lambda_j {}^t v_i v_j = \begin{cases} 0 & \text{si } i \neq j \\ \lambda_j & \text{si } i = j \end{cases}$$

et la matrice θ_G s'écrit par conséquent :

$$\theta_G = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

L'équation de l'ellipsoïde centré d'inertie de E s'écrit donc, par rapport aux axes $Gz_1, Gz_2, \dots, Gz_n$:

$$(z_1 \ z_2 \ \dots \ z_n) \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = 1 \quad ; \quad \text{donc :}$$

Proposition IO : L'équation de l'ellipsoïde centré d'inertie de E s'écrit, par rapport au système d'axes $Gz_1, Gz_2, \dots, Gz_n$:

$$\lambda_1 z_1^2 + \lambda_2 z_2^2 + \dots + \lambda_n z_n^2 = 1$$

C - LONGUEUR DES AXES DE L'ELLIPSOÏDE CENTRE D'INERTIE

Chacun des axes $Gz_1, Gz_2, \dots, Gz_n$ de l'ellipsoïde centré d'inertie coupe celui-ci en deux points, symétriques par rapport à G. Désignons par P_j et P_j^1 les points où l'axe Gz_j coupe la surface de l'ellipsoïde.

La longueur du segment $P_j^1 P_j$ est appelée la longueur de l'axe Gz_j de l'ellipsoïde. Nous nous proposons de calculer la longueur de chacun des axes.

Les points P_j^1 et P_j se trouvant tous les deux sur l'axe Gz_j et sur la surface de l'ellipsoïde, ont des coordonnées $z_1, z_2, \dots, z_n$ qui vérifient (voir proposition IO) les n relations :

$$\begin{cases} z_1 = 0 ; z_2 = 0 ; \dots ; z_{j-1} = 0 ; z_{j+1} = 0 ; \dots ; z_n = 0 \\ \lambda_1 z_1^2 + \lambda_2 z_2^2 + \dots + \lambda_n z_n^2 = 1 \end{cases}$$

c'est-à-dire encore :

$$\begin{cases} z_1 = 0 ; z_2 = 0 ; \dots ; z_{j-1} = 0 ; z_{j+1} = 0 ; \dots ; z_n = 0 \\ \lambda_j z_j^2 = 1 \end{cases}$$

ou

$$\begin{cases} z_i = 0 & (i \neq j) \\ z_j = \pm \sqrt{\frac{1}{\lambda_j}} \end{cases}$$

Les coordonnées de P_j et de P'_j , par rapport au système de référence $Gz_1, Gz_2, \dots, Gz_n$, sont donc :

$$\left\{ \begin{array}{l} P_j \quad (0, 0, \dots, 0, \sqrt{\frac{1}{\lambda_j}}, 0, \dots, 0) \\ P'_j \quad (0, 0, \dots, 0, -\sqrt{\frac{1}{\lambda_j}}, 0, \dots, 0) \end{array} \right. \quad (\text{voir figure 4})$$

La longueur de GP_j est donc $\sqrt{\frac{1}{\lambda_j}}$; celle de GP'_j est aussi égale à $\sqrt{\frac{1}{\lambda_j}}$; par conséquent :

Proposition 11 : La longueur de l'axe Gz_j de l'ellipsoïde est :

$$2 \sqrt{\frac{1}{\lambda_j}}$$

Ainsi, si les valeurs propres $\lambda_1, \lambda_2, \dots, \lambda_n$ de la matrice T_G sont ordonnées de telle sorte que :

$$0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

alors les axes $Gz_1, Gz_2, \dots, Gz_n$ sont ordonnés par longueurs décroissantes : Gz_1 est le plus long, Gz_n le plus court.

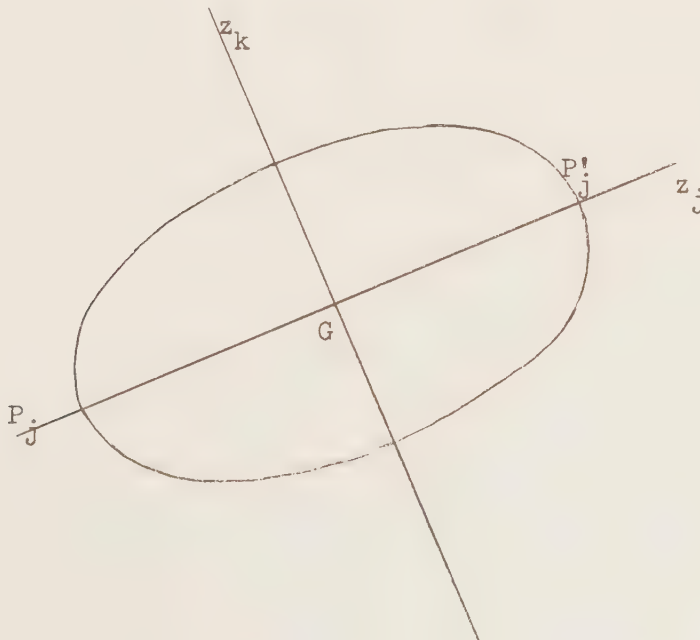


figure 4

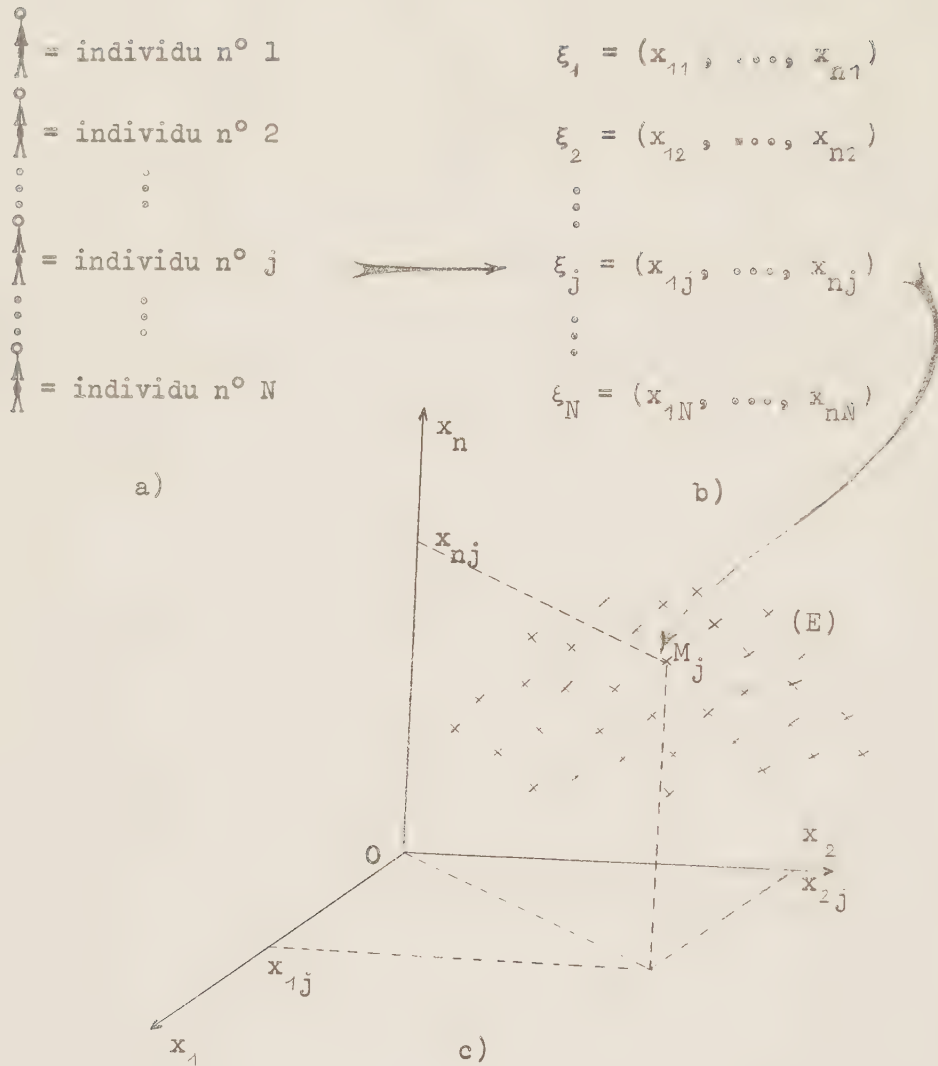


Figure I

Représentation géométrique de l'échantillon :

- en a), l'échantillon de taille N ;
- en b), à l'individu numéro j correspond la "valeur" $\xi_j = (x_{1j}, \dots, x_{nj})$ de la variable aléatoire X à n dimensions ;
- en c), à la valeur ξ_j (donc aussi au j-ième individu) correspond le point M_j de l'espace R^n , à l'ensemble des individus constituant l'échantillon correspond un ensemble E de points constituant le nuage statistique.

DEUXIEME PARTIE

INERTIE STATISTIQUE ET ANALYSE FACTORIELLE

CHAPITRE I : APERCU SUR L'ANALYSE FACTORIELLE

A - LE PROBLEME DE L'ANALYSE FACTORIELLE

Des branches très diverses de la Science s'intéressent à des ensembles de variables aléatoires simultanées, susceptibles d'avoir entre elles certaines relations a priori inconnues, mais devinées. Par exemple, la Biométrie considère plusieurs caractères mesurés sur un individu ou sur un organe ; la Psychométrie essaie de relier entre eux les résultats de plusieurs tests subis par une personne ; etc...

Dans un double but de simplification et d'explication, l'expérimentateur, postulant que "deux variables simultanées sont en liaison parce qu'elles dépendent toutes les deux d'un certain nombre de facteurs communs", cherchera à exprimer chacune des variables qu'il considère en fonction d'un petit nombre de "facteurs" communs à toutes ces variables. Mathématiquement, ce problème se traduit ainsi : étant donné n variables observées $x_1, x_2, \dots, x_n$, trouver k variables $z_1, z_2, \dots, z_k$ ($k < n$), et n fonctions $f_1, f_2, \dots, f_n$ de ces k variables, telles que l'on puisse écrire :

$$\begin{cases} x_1 = f_1(z_1, z_2, \dots, z_k) \\ x_2 = f_2(z_1, z_2, \dots, z_k) \\ \vdots \\ x_n = f_n(z_1, z_2, \dots, z_k) \end{cases}$$

L'ensemble des méthodes conçues pour tenter de résoudre ce problème constitue l'Analyse Factorielle. Le problème tel qu'il est posé est d'ailleurs insoluble ; il comporte plus d'inconnues que de données. Il est nécessaire d'ajouter à son énoncé certaines hypothèses, portant par exemple sur la distribution de probabilité des variables considérées, ou sur la forme des fonctions f_j . Ainsi, jusqu'à maintenant, on a cherché exclusivement des fonctions f_j , du type linéaire (en raison de la relative simplicité des calculs).

Le présent travail, dont le but est de présenter une méthode d'Analyse Factorielle, se trouvera lui aussi limité à la construction et à l'utilisation d'une méthode d'analyse factorielle linéaire. Le modèle mathématique recherché est alors :

$$\begin{cases} x_1 = a_{11} z_1 + a_{12} z_2 + \dots + a_{1k} z_k \\ x_2 = a_{21} z_1 + a_{22} z_2 + \dots + a_{2k} z_k \\ \vdots \\ x_n = a_{n1} z_1 + a_{n2} z_2 + \dots + a_{nk} z_k \end{cases}$$

B - APERÇU HISTORIQUE

Il semble que le problème de l'analyse factorielle a été posé pour la première fois (sous une forme simplifiée) au début du présent siècle. Il s'agissait d'une question de Psychologie : les différentes aptitudes d'un individu (aptitudes évaluées quantitativement au moyen de notes obtenues à des tests), sont-elles fonctions d'un seul facteur important ?

Ch. SPEARMAN, en 1904, a essayé de résoudre ce problème, en faisant l'hypothèse de l'existence d'un tel facteur (le facteur "g") et de la linéarité des fonctions reliant ce facteur aux aptitudes individuelles.

Cette solution a un seul facteur commun n'étant pas, dans la pratique, très satisfaisante, on a été amené à construire d'autres modèles mathématiques, supposant l'existence de plusieurs facteurs. Finalement THURSTONE, en 1930, a posé la question sous une autre forme : combien de facteurs communs sont nécessaires pour rendre compte des corrélations entre les variables observées ? C'était là, pratiquement, poser le problème sous sa forme actuelle.

L'interprétation vectorielle de l'échantillonnage, l'utilisation des techniques du calcul matriciel, ont permis l'élaboration de nombreuses méthodes d'analyse factorielle, sous l'impulsion de HOLZINGER, HOTELLING, THOMSON, THURSTONE, etc... Pour plus de détails sur ces questions, on pourra se référer à des ouvrages que l'on peut considérer comme des manuels d'analyse factorielle : ceux de HOLZINGER and HARMAN (1941), THURSTONE (1945), THOMSON (1939). La lecture des comptes-rendus du Colloque International sur l'Analyse Factorielle (C.N.R.S., 1955), sera également profitable.

Une phrase de C. BURT (1955) résume bien l'évolution des techniques d'analyse factorielle depuis les années 1900 : "un demi-siècle s'est écoulé depuis l'introduction des procédés factoriels en psychologie ; et pendant cette période, des méthodes de travail en nombre étonnant ont été proposées, dont chacune a sa chapelle".

Bien entendu, outre le perfectionnement de ses techniques, l'analyse factorielle a vu, au cours de cette période, son champ d'application s'élargir, et passer de la seule Psychologie à l'ensemble des Sciences expérimentales. En Biologie par exemple, nous citerons LEDERMAN (1955), MATSAKIS (1961), SCHREIDER (1955), PEARCE et HOLLAND (1961).

C - METHODE DES COMPOSANTES PRINCIPALES

Cette méthode, mise au point par HOTELLING (1933 et 1936), mais déjà annoncée par les travaux de PEARSON (1901), nous semble être la meilleure des méthodes utilisées jusqu'ici.

On envisage une variable aléatoire à n dimensions, soit

$$X = (X_1, X_2, \dots, X_n)$$

Si l'on suppose que X est une variable normale, la densité de probabilité

en un point $x = (x_1, x_2, \dots, x_n)$ est de la forme

$$p(x) = K e^{-\frac{1}{2} Q(x_1, x_2, \dots, x_n)}$$

Dans cette expression, $Q(x_1, x_2, \dots, x_n)$ désigne la forme quadratique

$${}^t(x - \mu) S^{-1} (x - \mu)$$

S^{-1} est la matrice inverse de la matrice des covariances de $X = (X_1, X_2, \dots, X_n)$; μ est le vecteur moyen, de composantes $\mu_1, \mu_2, \dots, \mu_n$ (μ_j = moyenne de x_j) ; de sorte que :

$$Q(x_1, x_2, \dots, x_n) = (x_1 - \mu_1 \dots x_n - \mu_n) \begin{bmatrix} \vdots & & \\ & S^{-1} & \\ \vdots & & \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_n - \mu_n \end{bmatrix}$$

K est un coefficient numérique, choisi de telle sorte que

$$\int_{R^n} p(x) dx = 1$$

(pour plus de détails sur ces questions, voir ANDERSON, 1958, chapitre 2).

Pour une telle variable aléatoire normale à n dimensions, les surfaces d'égale densité de probabilité (dans l'espace R^n muni du système de référence $Ox_1, Ox_2, \dots, Ox_n$) sont obtenues en écrivant que :

$$p(x) = \text{constante, c'est-à-dire :}$$

$$K e^{-\frac{1}{2} Q(x_1, x_2, \dots, x_n)} = \text{constante,}$$

soit finalement : $Q(x_1, x_2, \dots, x_n) = \text{constante}$.

Cette dernière relation, liant les coordonnées $x_1, x_2, \dots, x_n$ du point x , est l'équation d'une conique à n dimensions, et on démontre qu'il s'agit d'un ellipsoïde (voir ANDERSON, ouvrage cité).

Les "composantes principales", au sens d'HOTELLING, sont les composantes de X suivant les axes de cet ellipsoïde, c'est-à-dire suivant les directions propres de la matrice carrée symétrique S^{-1} , lesquelles directions propres coïncident d'ailleurs avec celles de S . On obtient donc autant de composantes principales qu'il y a de variables numériques (factorisation complète). En fait, on sélectionne, parmi ces n facteurs communs que sont les n composantes principales, ceux qui apportent la contribution la plus importante à la variation des variables $X_1, X_2, \dots, X_n$. Cette sélection se fait au moyen d'une discussion sur les valeurs des variances de ces facteurs : un facteur est jugé d'autant plus important que sa variance est plus grande : or ces variances peuvent être connues, ce sont les valeurs propres de la matrice S (ou encore : les inverses des valeurs propres de la matrice S^{-1}).

La "factorisation" par la méthode d'HOTELLING revient donc à la recherche des valeurs propres et directions propres de la matrice S , et à une discussion sur les valeurs relatives de ces valeurs propres.

D - COVARIANCES ET CORRELATIONS

Soit X une variable aléatoire de moyenne μ , d'écart-type σ . La "standardisation" de X est le remplacement de la variable X par la variable

$$Y = \frac{X - \mu}{\sigma}$$

La variable aléatoire Y a alors pour moyenne 0 et pour écart-type 1.

Le coefficient de corrélation de deux variables X_1 et X_2 , soit :

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

peut alors être considéré comme la covariance des deux variables standardisées correspondantes :

$$Y_1 = \frac{X_1 - \mu_1}{\sigma_1} \quad Y_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

Signalons que beaucoup d'utilisateurs des différentes méthodes

d'Analyse Factorielle, préfèrent travailler sur la matrice R des corrélations, plutôt que sur la matrice S des covariances. Ceci revient donc à travailler sur la matrice des covariances des variables standardisées. Ce faisant, ces utilisateurs estiment que les variables brutes ne sont pas "comparables" (au sens de l'Analyse Factorielle), et que la standardisation les rend comparables.

Nous ne discuterons pas ici ce point de vue. Nous constaterons simplement qu'une matrice de corrélations "contient moins d'information" que la matrice de covariances dont elle découle. En effet, d'une part, tous les renseignements que nous pouvons tirer de la matrice des corrélations, sont déjà contenus en puissance dans la matrice des covariances ; d'autre part, plusieurs matrices de covariances distinctes peuvent engendrer une même matrice des corrélations.

Un exemple (à deux dimensions) : soient deux variables aléatoires X_1 et X_2 ; les matrices S et R sont respectivement

$$S = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix} ; \quad R = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}$$

avec $s_1^2 = \text{var}(X_1)$, $s_2^2 = \text{var}(X_2)$, $s_{12} = s_{21} = \text{cov}(X_1, X_2)$

$$r_{12} = r_{21} = r(X_1, X_2) = \frac{s_{12}}{s_1 s_2}$$

Si k_1 et k_2 sont des nombres positifs quelconques, les deux variables aléatoires $X'_1 = k_1 X_1$ et $X'_2 = k_2 X_2$ sont telles que :

$$S' = \begin{pmatrix} k_1^2 s_1^2 & k_1 k_2 s_{12} \\ k_2 k_1 s_{21} & k_2^2 s_2^2 \end{pmatrix} ; \quad R' = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} = R$$

On a bien, en effet :

$$r(X'_1, X'_2) = \frac{k_1 k_2 s_{12}}{k_1 s_1 k_2 s_2} = \frac{s_{12}}{s_1 s_2} = r(X_1, X_2)$$

CHAPITRE II : UTILISATION DE L'INERTIE STATISTIQUE

A - UN CHANGEMENT DE SYSTEME DE REFERENCE. INTERPRETATION STATISTIQUE

Dans l'espace R^n (où l'échantillon considéré est représenté par un ensemble de N points), le système de référence initialement choisi était $Ox_1, Ox_2, \dots, Ox_n$ (voir signification au début de la Première Partie).

Choisissons maintenant, dans ce même espace R^n , un autre système de référence, formé par les n axes centraux d'inertie de E , soit $Gz_1, Gz_2, \dots, Gz_n$, normés.

Ainsi un point M de l'espace, qui avait par rapport à l'ancien système des coordonnées $x_1, x_2, \dots, x_n$, a maintenant pour coordonnées, par rapport au nouveau système, des nombres $z_1, z_2, \dots, z_n$. Il existe entre les x_j d'une part, les z_k d'autre part, des relations résumées par les égalités matricielles :

$$X = C Z \quad ; \quad Z = {}^t C X \quad (\text{avec } {}^t C = C^{-1})$$

(La deuxième égalité notamment permet le calcul pratique des z_k connaissant les x_j et la matrice ${}^t C$).

En particulier, les N points M_i formant le nuage statistique E , ont eux aussi des coordonnées z_j par rapport au nouveau système d'axes : le point M_i , dont les coordonnées par rapport à l'ancien système étaient $x_{1i}, x_{2i}, \dots, x_{ni}$, a par rapport au nouveau système des coordonnées $z_{1i}, z_{2i}, \dots, z_{ni}$, telles que

$$z_{ji} = a_{j1} (x_{1i} - \bar{x}_1) + a_{j2} (x_{2i} - \bar{x}_2) + \dots + a_{jn} (x_{ni} - \bar{x}_n)$$

On définit ainsi, pour les individus de la population dont on a étudié un échantillon, n nouvelles variables aléatoires :

$$Z_j = a_{j1} (X_1 - \bar{x}_1) + \dots + a_{jn} (X_n - \bar{x}_n) \quad j = 1, 2, \dots, n$$

Chacune de ces variables Z_j étant attachée à la population étudiée, on peut dire qu'il s'agit d'un caractère des individus composant cette population. Mais, alors que les X_k étaient des caractères qu'on pouvait supposer directement mesurables, les Z_j sont des caractères non direc-

tement mesurables, mais simplement calculables dès que les valeurs numériques des X_k sont connues.

L'existence de caractères non directement mesurables n'a d'ailleurs rien d'extraordinaire ; c'est même un cas assez général : il n'est pas rare d'évaluer des poids ou des volumes, par exemple, à partir de dimensions elles-mêmes mesurées directement. D'autre part, même les cas où les Z_j n'auraient pas d'interprétation matérielle ne seraient pas non plus extraordinaires : le coefficient de robusticité, l'indice céphalique sont des notions familières en Biométrie humaine ; le quotient intellectuel n'est pas moins familier au Psychologue.

Et pourtant, ces quantités, fonctions elles-mêmes de caractères mesurables, n'ont pas de support matériel chez l'individu ; cependant, chacune caractérise un aspect du comportement global de l'individu.

Ainsi en est-il des variables Z_j : chacune intéresse globalement l'individu (ou l'organe) sur lequel sont mesurés les X_k . Leur signification, leur utilisation pratique sera affaire d'interprétation dans chacun des cas numériques envisagés. Nous reviendrons plus loin sur cette question.

B - CALCUL DES MOMENTS DES VARIABLES Z_j , DANS L'ECHANTILLON E

a) Les moyennes :

$$z_{ji} = a_{j1}(x_{1i} - \bar{x}_1) + a_{j2}(x_{2i} - \bar{x}_2) + \dots + a_{jn}(x_{ni} - \bar{x}_n)$$

$$\sum_{i=1}^N z_{ji} = a_{j1} \sum_{i=1}^N (x_{1i} - \bar{x}_1) + \dots + a_{jn} \sum_{i=1}^N (x_{ni} - \bar{x}_n)$$

$$\text{or } \sum_{i=1}^N (x_{ki} - \bar{x}_k) = 0 ,$$

par définition de $\bar{x}_k$ comme moyenne de X_k dans l'échantillon. Donc :

$$\sum_{i=1}^N z_{ji} = 0 , \text{ et par suite } \bar{z}_j = 0$$

b) Les variances

$$\text{var}(z_j) = \frac{1}{N} \sum_{i=1}^N (z_{ji} - \bar{z}_j)^2 = \frac{1}{N} \sum_{i=1}^N z_{ji}^2 \quad ; \quad \text{mais :}$$

$$z_{ji}^2 = a_{j1}^2 (x_{1i} - \bar{x}_1)^2 + \dots + a_{jn}^2 (x_{ni} - \bar{x}_n)^2 + 2 a_{j1} a_{j2} (x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2) + \dots$$

$$\text{d'où : } \text{var}(z_j) = a_{j1}^2 s_1^2 + \dots + a_{jn}^2 s_n^2 + 2 a_{j1} a_{j2} s_{12} + \dots$$

c) Les covariances :

$$\text{cov}(z_j, z_k) = \frac{1}{N} \sum_{i=1}^N (z_{ji} - \bar{z}_j)(z_{ki} - \bar{z}_k) = \frac{1}{N} \sum_{i=1}^N z_{ji} z_{ki}$$

$$z_{ji} z_{ki} = a_{j1} a_{k1} (x_{1i} - \bar{x}_1)^2 + \dots + a_{jn} a_{kn} (x_{ni} - \bar{x}_n)^2 + a_{j1} a_{k2} (x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2) + a_{j2} a_{k1} (x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2) + \dots \quad ; \quad \text{d'où :}$$

$$\text{cov}(z_j, z_k) = a_{j1} a_{k1} s_1^2 + \dots + a_{jn} a_{kn} s_n^2 + (a_{j1} a_{k2} + a_{j2} a_{k1}) s_{12} + \dots$$

On peut résumer le calcul des variances et covariances sous la forme matricielle :

$$V = \begin{bmatrix} \text{var}(z_1) & \text{cov}(z_1, z_2) & \dots & \text{cov}(z_1, z_n) \\ \text{cov}(z_2, z_1) & \text{var}(z_2) & \dots & \text{cov}(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(z_n, z_1) & \text{cov}(z_n, z_2) & \dots & \text{var}(z_n) \end{bmatrix} = {}^t C S C,$$

où les matrices C et S ont les significations données plus haut, soit :

$$C = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & & a_{nn} \end{bmatrix} \quad \text{et } S = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1n} \\ s_{21} & s_2^2 & & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & & s_n^2 \end{bmatrix}$$

Or, rappelons que $S = s_E^2 I - T_G$, d'où :

$$t_{CSC} = s_E^2 t_{CIC} - t_{CT_G C}$$

$$= s_E^2 I - \theta_G$$

avec $t_{CIC} = t_{CC} = I$ (car $t_C = C^{-1}$), et :

$$\theta_G = t_{CT_G C} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

$$\text{Donc : } t_{CSC} = \begin{pmatrix} s_E^2 - \lambda_1 & 0 & \dots & 0 \\ 0 & s_E^2 - \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_E^2 - \lambda_n \end{pmatrix}$$

En conclusion, nous pouvons énoncer :

Proposition I2 : Les moyennes, variances et covariances d'échantillon des variables Z_j , dans l'échantillon considéré, sont :

$$\bar{z}_j = 0 ; \text{ var } (z_j) = s_E^2 - \lambda_j \text{ (noté } \lambda_j') ; \text{ cov } (z_j, z_k) = 0 \text{ (} j \neq k \text{)}$$

Ceci signifie, en particulier, que les variables Z_j sont deux à deux incorréllées (dans l'échantillon).

C - EXPRESSION DE LA MATRICE S EN FONCTION DES λ_j'

Nous avons ci-dessus exprimé la matrice V des covariances des z_j , en fonction de la matrice S des covariances des x_k : $V = t_{CSC}$.

Réciproquement, nous pouvons exprimer S en fonction de V :

$$S = t_C^{-1} t_{CSC}^{-1} = t_C^{-1} V C^{-1} = C V t_C$$

(rappelons en effet que $C^{-1} = {}^t C$, et par conséquent ${}^t C^{-1} = C$). Donc :

Proposition I3 : on a la relation $S = CV^t C$ (S matrice des covariances des x_k , V matrice des covariances des z_j , dans l'échantillon) ; c'est-à-dire encore :

$$\left\{ \begin{array}{l} s_j^2 = a_{1j}^2 \lambda_1' + a_{2j}^2 \lambda_2' + \dots + a_{nj}^2 \lambda_n' \\ s_{jk} = a_{1j} a_{1k} \lambda_1' + a_{2j} a_{2k} \lambda_2' + \dots + a_{nj} a_{nk} \lambda_n' \end{array} \right.$$

Remarque : il sera utile par la suite de connaître l'expression de s_E^2 en fonction des λ_j' . On a :

$$\begin{aligned} s_E^2 &= s_1^2 + s_2^2 + \dots + s_n^2 = \sum_{j=1}^N s_j^2 \quad (\text{par définition}) \\ &= \sum_{j=1}^N (a_{1j}^2 \lambda_1' + a_{2j}^2 \lambda_2' + \dots + a_{nj}^2 \lambda_n') \quad (\text{proposition I3}) \\ &= \lambda_1' \sum_{j=1}^N a_{1j}^2 + \dots + \lambda_n' \sum_{j=1}^N a_{nj}^2 ; \end{aligned}$$

mais rappelons que $a_{k1}, a_{k2}, \dots, a_{kn}$ sont les composantes du vecteur $\vec{v}_k$, vecteur propre unitaire associé à la valeur propre λ_k . Donc :

$$\sum_{j=1}^N a_{kj}^2 = 1 \quad \text{et par conséquent : } s_E^2 = \lambda_1' + \lambda_2' + \dots + \lambda_n'$$

Proposition I4 : on a la relation

$$s_E^2 = s_1^2 + s_2^2 + \dots + s_n^2 = \lambda_1' + \lambda_2' + \dots + \lambda_n'$$

Ce résultat pouvait se démontrer également en faisant appel à la propriété suivante, relative aux matrices carrées : la somme des valeurs propres d'une matrice carrée est égale à la trace de cette matrice (c'est-à-dire : à la somme de ses éléments diagonaux). Il suit de là en effet que :

$$\lambda_1 + \lambda_2 + \dots + \lambda_n = \text{trace de } T_G = (s_E^2 - s_1^2) + \dots + (s_E^2 - s_n^2) \\ = (n - 1) s_E^2 \quad ; \text{ et d'autre part :}$$

$$\lambda_1' + \lambda_2' + \dots + \lambda_n' = (s_E^2 - \lambda_1) + \dots + (s_E^2 - \lambda_n) \quad (\text{proposition I2}) \\ = n s_E^2 - (\lambda_1 + \lambda_2 + \dots + \lambda_n)$$

$$\text{Donc : } \lambda_1' + \lambda_2' + \dots + \lambda_n' = n s_E^2 - (n - 1) s_E^2 = s_E^2$$

D - APPLICATION A L'ANALYSE FACTORIELLE

L'expression des variables $X_1, X_2, \dots, X_n$ comme fonctions linéaires d'autres variables $Z_1, Z_2, \dots, Z_n$ (incorrélées deux à deux), constitue une réponse possible au problème de l'analyse factorielle de l'ensemble des variables $X_1, X_2, \dots, X_n$. Ainsi les variables $Z_1, Z_2, \dots, Z_n$ peuvent être considérées comme des "facteurs communs", au sens de l'analyse factorielle.

En fait, il s'agit là d'une "factorisation complète" : nous nous trouvons en présence d'autant de facteurs qu'il y a de variables, et chacune des variables est exprimable exactement comme fonction linéaire des facteurs, sous la forme :

$$X_j - \bar{x}_j = a_{1j} Z_1 + a_{2j} Z_2 + \dots + a_{nj} Z_n$$

Cette solution ne nous satisfait donc pas entièrement, puisque notre but est de trouver moins de facteurs communs que de variables (le nombre de facteurs communs devant être "le plus petit possible"). Il nous faut donc envisager la suppression de certains des facteurs Z_j .

Préalablement à cette suppression, nous devons dégager la notion d'"importance" d'un facteur. En effet, intuitivement, nous pouvons penser que ces n facteurs, bien que tous nécessaires à l'expression exacte de chaque variable, ne sont pas aussi importants les uns que les autres : la suppression de certains d'entre eux peut ne pas changer grand-chose aux valeurs calculées des variables qui, sans être rigoureusement égales aux

valeurs observées, en seront cependant suffisamment voisines pour qu'on puisse raisonnablement les confondre.

Pour la commodité de l'exposé, nous appellerons variance globale de l'ensemble de variables $(X_1, X_2, \dots, X_n)$ (dans l'échantillon considéré) la quantité

$$s_E^2 = s_1^2 + s_2^2 + \dots + s_n^2$$

Ainsi, la proposition I4 exprime le fait que la variance globale de l'ensemble de variables $(X_1, X_2, \dots, X_n)$ est égale à la variance globale de l'ensemble de facteurs $(Z_1, Z_2, \dots, Z_n)$ (autrement dit, cette factorisation conserve la variance globale).

Revenons à la suppression des facteurs peu importants. Il nous faut, bien entendu, donner une définition mathématique de l'"importance" d'un facteur, afin de pouvoir calculer cette importance, et la comparer à celles des autres facteurs.

Il serait trop long de construire ici une axiomatique de l'importance des facteurs. Nous nous contenterons de rechercher une certaine quantité numérique, permettant d'effectuer des raisonnements semblables à ceux que nous permet la notion d'importance, dans le sens commun du terme.

Or, la variance d'un facteur correspond assez bien à la notion d'importance. En effet, d'après la proposition I2, nous savons que :

$$\begin{cases} \text{moyenne de } Z_k = 0 \\ \text{variance de } Z_k = s_E^2 - \lambda_k = \lambda_k' \end{cases}$$

L'inégalité de Tchébycheff s'écrit donc, pour Z_k :

$$\text{Prob} (| Z_k | > t \sqrt{\lambda_k'}) < \frac{1}{t^2} \quad (t \text{ nombre positif quelconque})$$

On peut en conclure en particulier que plus la variance λ_k' est petite, moins le facteur Z_k a de chances de prendre des valeurs éloignées de la valeur zéro.

Nous pouvons donc choisir comme mesure de l'importance d'un facteur une fonction directe de sa variance.

Une première solution consiste à choisir comme mesure de l'importance de Z_k sa variance elle-même, soit λ_k' . Signalons que cette solution

est souvent utilisée dans les méthodes d'analyse factorielle où l'on a besoin d'éliminer les facteurs les moins importants (par exemple : méthode d'HOTELLING). (voir : PEARCE and HOLLAND, 1961).

La proposition 14 nous indique que les importances des n facteurs Z_k sont additives, leur somme étant la variance globale s_E^2 . Pour cette raison, la variance λ_k sera appelée l'importance globale de Z_k . Comme il s'agit en fait de comparer entre eux les facteurs, du point de vue de leur importance, nous utiliserons aussi l'importance globale relative, soit

$$\frac{\lambda_k}{s_E^2} \quad (\text{exprimée par exemple en pourcentage}).$$

Une fois connues les importances globales relatives des différents facteurs, on supprime le plus possible de facteurs de faible importance, les facteurs restants devant néanmoins avoir pour importance totale une part très large de la variance globale (part laissée au jugement de l'opérateur).

Pour illustrer cette méthode de suppression des facteurs les moins importants, supposons un cas où $n = 4$, avec une variance globale $s_E^2 = 200$; les quatre facteurs communs ayant pour variances respectivement 150, 40, 6, 4, c'est-à-dire que les importances globales relatives sont 75 %, 20 %, 3 %, 2 %. Les deux premiers facteurs ont une importance totale de 95 %. Il n'est donc pas déraisonnable de négliger les deux autres facteurs, et de considérer que les 4 variables dépendent de deux facteurs communs.

Cette première solution, pour supprimer les facteurs les moins importants, ne nous satisfait pas. En effet, un certain facteur, dont l'importance globale est jugée négligeable par cette méthode, peut se révéler, par le jeu des coefficients a_{ij} , particulièrement important pour telle ou telle variable X_j . C'est pourquoi nous préférons utiliser, dans la suite, un deuxième type de solution, que nous allons exposer ci-dessous.

La "factorisation complète", c'est-à-dire l'expression exacte de chaque variable X_j en fonction des n facteurs Z_k , nous a permis d'établir les trois relations suivantes :

$$\left\{ \begin{array}{l} X_j - \bar{x}_j = a_{1j}Z_1 + a_{2j}Z_2 + \dots + a_{nj}Z_n \quad (\text{prop. 9}) \\ s_j^2 = a_{1j}^2 \lambda_1' + a_{2j}^2 \lambda_2' + \dots + a_{nj}^2 \lambda_n' \quad (\text{prop. I3}) \\ s_{jk} = a_{1j}a_{1k} \lambda_1' + \dots + a_{nj}a_{nk} \lambda_n' \quad (\text{prop. I3}) \end{array} \right.$$

Dans la première de ces trois relations, le terme $a_{kj}Z_k$ représente l'apport du facteur Z_k à la quantité $X_j - \bar{x}_j$. En première analyse, on pourra dire que le facteur Z_k est d'autant moins important (du point de vue de X_j) que ce terme $a_{kj}Z_k$ est petit.

Sauf cas particulier, le terme $a_{kj}Z_k$ ne sera jamais identiquement nul, dans la pratique. Il s'agit d'une variable aléatoire dont nous connaissons la moyenne et la variance, d'après la proposition I2 :

$$\left\{ \begin{array}{l} \text{moy} (a_{kj}Z_k) = a_{kj} \quad \text{moy} (Z_k) = 0 \\ \text{var} (a_{kj}Z_k) = a_{kj}^2 \text{var} (Z_k) = a_{kj}^2 \lambda_k' \end{array} \right.$$

L'inégalité de Tchébycheff, appliquée au terme $a_{kj}Z_k$, s'écrit donc :

$$\text{Prob} (| a_{kj}Z_k | > t \sqrt{a_{kj}^2 \lambda_k'}) < \frac{1}{t^2}$$

Comme plus haut au sujet du facteur Z_k , nous concluons ici que plus la variance $a_{kj}^2 \lambda_k'$ est petite, moins le terme $a_{kj}Z_k$ a des chances de prendre des valeurs éloignées de zéro.

Nous prendrons donc comme mesure de l'importance du facteur Z_k pour la variable X_j , la quantité $a_{kj}^2 \lambda_k'$. Remarquons que cette quantité représente également la contribution de Z_k à la variance de X_j , puisque :

$$s_j^2 = a_{1j}^2 \lambda_1' + a_{2j}^2 \lambda_2' + \dots + a_{nj}^2 \lambda_n'$$

Là encore, les importances des n facteurs apparaissent comme additives, leur somme étant la variance s_j^2 . Autre interprétation : la variance s_j^2 est une somme pondérée des importances globales λ_k' ; ce sont précisément les coefficients de pondération, du type a_{kj}^2 , qui nous permettent

de nuancer l'importance de chaque facteur Z_k , suivant la variable X_j considérée.

Donc : $a_{kj}^2 \lambda_k$ sera appelé l'importance de Z_k pour la variable X_j ; pour rendre les comparaisons plus aisées, nous utiliserons surtout l'importance relative de Z_k pour la variable X_j , soit :

$$\frac{a_{kj}^2 \lambda_k}{s_j^2}$$

Une fois connues les n importances relatives des Z_k pour la même variable X_j , nous pouvons comparer entre elles ces importances ; nous supprimons le plus possible de facteurs de faible importance, étant entendu que les facteurs restants doivent apporter à la variance s_j^2 une contribution totale jugée presque égale à cette variance elle-même. A ce propos, la connaissance de la part de s_j^2 provenant des erreurs de mesure ou d'échantillonnage, donnera plus de puissance à la méthode ; en effet, ce serait un excellent argument de supprimer les facteurs dont la contribution totale à s_j^2 est de l'ordre de grandeur de ces "erreurs", et de ne supprimer que ces facteurs-là. Malheureusement, il n'est souvent pas possible d'évaluer ces erreurs.

A la fin des discussions portant sur chaque variable X_j , nous disposons, globalement, d'au plus n facteurs $Z_1, Z_2, \dots, Z_n$; mais, pour chaque X_j , le nombre de facteurs intervenant dans la variation de $X_j - \bar{x}_j$ est moindre que n . Un même facteur Z_k peut être commun à toutes les variables X_j ; il peut également n'apparaître en fin de compte que pour l'une des variables, ou pour aucune ; entre ces deux cas extrêmes, il peut être commun à plusieurs variables mais pas à toutes, ce qui permettra de l'interpréter.

E - RETOUR A LA METHODE D'HOTELLING

La Théorie de l'Inertie Statistique, exposée dans la première partie de ce travail, nous a permis de construire la méthode d'analyse factorielle que nous venons de traiter ci-dessus.

D'autre part, nous avons donné plus haut un aperçu de la méthode

des composantes principales d' HOTELLING. Il est visible que notre méthode, et celle d' HOTELLING, ne sont pas fondamentalement différentes : dans les deux cas, l'analyse consiste :

1) à déterminer les valeurs propres et les vecteurs propres d'une matrice obtenue à partir de la matrice S des covariances ;

2) à prendre comme "facteurs communs" aux variables considérées, les composantes de ces variables suivant ces vecteurs propres ;

3) à éliminer, parmi ces facteurs, ceux qui sont jugés "de moindre importance" (élimination faite au moyen d'une discussion sur les variances de ces facteurs).

La seule différence majeure entre les deux méthodes est que, pour la méthode d' HOTELLING, les variables doivent être supposées gaussiennes dans leur ensemble, alors que pour notre méthode, aucune hypothèse n'est à faire au sujet de la distribution des variables.

En fait, cette différence n'est qu'apparente ; c'est une question d'interprétation. HOTELLING a besoin de supposer la distribution gaussienne parce qu'alors les axes principaux s'interprètent comme étant les axes des ellipsoïdes d'égale densité de probabilité. Dans notre méthode, nous interprétons les axes principaux comme étant les axes de l'ellipsoïde d'inertie (sans avoir besoin de supposer qu'il existe ou non des ellipsoïdes d'égale densité, donc de supposer que la distribution est ou non gaussienne).

Si la distribution considérée est effectivement gaussienne, alors notre méthode et celle d' HOTELLING coïncident : en effet, l'ellipsoïde d'inertie a les mêmes axes que les ellipsoïdes d'égale densité de probabilité ; cela tient au fait que les deux matrices S et T_G ont (dans ce cas comme dans tout autre cas d'ailleurs) les mêmes directions propres (cf Appendice). Reprenons d'ailleurs le parallèle entre la Mécanique Rationnelle et le Calcul des Probabilités : en Mécanique, on définit l'ellipsoïde centré d'inertie de tout corps solide, quelle que soit sa forme ; en particulier, un corps solide de forme ellipsoïdale (et dont les surfaces d'égale densité sont des ellipsoïdes semblables) admet pour ellipsoïde centré d'inertie un ellipsoïde homothétique du solide lui-même, donc ayant mêmes axes.

Il ressort de ce qui précède que notre méthode est une généralisation de la méthode d' HOTELLING. Comme dans la pratique, il s'avère que

les deux méthodes sont identiques dans leur déroulement, il serait peut-être plus précis de dire que nous avons non pas généralisé la méthode d'HOTELLING, mais étendu son domaine d'application, en montrant qu'elle est valable pour toute distribution, gaussienne ou non (pourvu qu'existent ses moments d'ordre un et deux), sous réserve d'interpréter les axes principaux non comme axes d'ellipsoïdes d'égale densité, mais comme axes de l'ellipsoïde centré d'inertie.

C H A P I T R E I I I : A N A L Y S E F A C T O R I E L L E E T B I O M E T R I E

A - I N T E R P R E T A T I O N D E L ' A N A L Y S E F A C T O R I E L L E E T D E S E S R E S U L T A T S

Nous avons vu ci-dessus qu'il existe des méthodes (en particulier celle que nous avons construite), permettant d'exprimer (approximativement) n variables aléatoires simultanées comme fonctions linéaires de "facteurs communs", en nombre plus petit. Mais l'existence d'une possibilité technique ne saurait, à elle seule, justifier que l'on procède systématiquement à une analyse factorielle sur n importe quel lot de variables aléatoires simultanées. Il faut encore que les variables considérées soient telles que l'existence de facteurs communs puisse être raisonnablement admise, et que l'on puisse trouver a posteriori une signification pratique aux facteurs déterminés par le calcul.

C'est le cas en Biométrie, comme nous allons le voir plus loin. Mais bien entendu, il ne faut pas s'attendre à trouver un support matériel bien défini aux facteurs trouvés. Ainsi, nous avons dit plus haut que les facteurs calculés étaient des entités numériques analogues, par la manière dont ils sont construits, à des coefficients tels que l'indice céphalique ou le coefficient de robusticité. Il faut pousser plus loin l'analogie, et considérer que les facteurs calculés, pour être valables, doivent signifier quelque chose pour l'organisme, même et surtout s'il s'agit d'une signification d'ensemble, concernant l'organisme globalement. C'est par exemple le

cas pour l'indice céphalique, qui est une "mesure" de la forme du crâne ; pour le coefficient de robusticité, qui est une "mesure" de la robustesse du corps.

Nous allons voir ci-dessous comment on peut justifier a priori la pratique d'une analyse factorielle sur des variables aléatoires qui sont des caractères biologiques. Puis, dans la troisième partie de ce travail, nous essaierons de donner une signification a posteriori aux facteurs que nous déterminerons par l'application de l'analyse factorielle dans des cas numériques précis.

B - LES HYPOTHESES FONDAMENTALES

La première hypothèse, celle qui va justifier nos essais d'Analyse Factorielle en Biologie, est l'hypothèse déterministe, que nous pouvons énoncer :

Hypothèse 1 : "tous les caractères d'un même individu sont complètement déterminés par un certain nombre de caractères fondamentaux, appelés facteurs déterminants (ou plus simplement : facteurs)".

Cette hypothèse n'est pas gratuite : si nous pensons que la structure, le fonctionnement, le comportement d'un organisme sont complètement déterminés par la structure génique de cet organisme, et par le milieu dans lequel il vit, nous pensons également à relier ces facteurs hypothétiques au complexe génotype-milieu. La liaison pouvant exister entre facteurs déterminants et complexe génotype-milieu nous est d'ailleurs inconnue, mais il est loisible de supposer qu'elle est étudiable.

Précisons davantage l'hypothèse déterministe, en disant que, pour un certain type d'organisme, elle fait état de k facteurs $z_1, z_2, \dots, z_k$, facteurs qui se retrouvent dans tous les organismes semblables, avec des valeurs différentes ou non. Il n'est pas exclu que, pour un même individu, les valeurs de ces facteurs puissent varier. Au contraire, puisqu'il est admis que les facteurs déterminants peuvent être reliés au milieu (et non pas seulement au génotype), nous pouvons penser que des variations des conditions de milieu doivent entraîner des variations des facteurs.

Dans l'incapacité (que nous espérons provisoire) de préciser da-

vantage la nature exacte de ces facteurs hypothétiques, ainsi que le genre de relations liant facteurs et caractères, nous sommes obligés d'émettre une autre hypothèse, destinée à axer notre champ d'investigation. Cette hypothèse est celle du déterminisme fonctionnel :

Hypothèse 2 : "la liaison existant entre un caractère x et les facteurs $z_1, z_2, \dots, z_k$ déterminant la structure de l'organisme, est une relation fonctionnelle".

Ceci nous amène à écrire :

$$x = f_x (z_1, z_2, \dots, z_k)$$

Par leur nature, les facteurs $z_1, z_2, \dots, z_k$ sont caractéristiques du type d'organisme (par exemple : de l'espèce) ; par leur valeur, ils sont caractéristiques de l'individu pris dans une population d'organismes de même type. La fonction f_x , elle, est caractéristique du caractère x : elle a la même forme quel que soit l'individu considéré dans la population étudiée.

Remarquons que, dans ces deux premières hypothèses, il n'est pas besoin de supposer que les caractères et les facteurs sont des variables numériques. Le mot "fonction" ou "relation fonctionnelle" doit être pris dans son sens mathématique le plus large. Les caractères et facteurs peuvent être "qualitatifs" et non quantitatifs ; leurs "valeurs" seront alors simplement les états qu'ils peuvent prendre.

L'hypothèse du déterminisme fonctionnel offrant un champ d'investigation encore trop vaste, nous devons restreindre ce champ, en fixant à l'avance la forme des fonctions que nous allons rechercher. D'où cette hypothèse de linéarité qui va nous permettre de commencer notre recherche.

Hypothèse 3 : "un caractère quantitatif x dépend de facteurs eux-mêmes quantitatifs. La relation fonctionnelle existant entre le caractère x et les facteurs $z_1, z_2, \dots, z_k$ est une relation linéaire".

Cette dernière hypothèse, à l'inverse des deux autres, n'apparaît pas très plausible. La linéarité a été choisie d'une manière tout à fait arbitraire. En fait, seule sa simplicité la justifie. Nous espérons d'une part que, pour le moment, les solutions linéaires au problème factoriel seront suffisantes (c'est-à-dire : elles constitueront une "bonne approxi-

mation" des fonctions réellement (?) mises en jeu) ; d'autre part, que des méthodes seront mises au point pour préciser la forme exacte de ces fonctions (ou pour en donner des approximations meilleures que les fonctions linéaires) (voir dans C.N.R.S., 1955, l'exposé de G. DARMOIS).

C - UNE VARIANTE : L'ANALYSE FACTORIELLE LOG-LINEAIRE

D'importants travaux, dus notamment à G. TEISSIER (1931, 1937), ont montré que, d'une manière assez générale, deux caractères quantitatifs x et y d'un même organisme sont liés par une relation du type $y = Kx^\alpha$ (où K et α sont deux constantes) ; c'est la relation d'allométrie, ou loi de dysharmonie.

Nous pouvons alors supposer que la relation liant un caractère x aux facteurs $z_1, z_2, \dots, z_k$ est du même type, soit :

$$x = K z_1^{\alpha_1} z_2^{\alpha_2} \dots z_k^{\alpha_k}$$

Remarquons qu'une telle relation est équivalente à la suivante :

$$\log x = \log K + \alpha_1 \log z_1 + \dots + \alpha_k \log z_k$$

Il existerait donc, dans cette hypothèse, une relation linéaire non pas entre chaque caractère et les facteurs déterminants, mais entre le logarithme de ce caractère et les logarithmes de ces facteurs.

Si nous considérons que l'allométrie a fait l'objet de recherches très sérieuses, aussi bien du point de vue descriptif (les vérifications de la loi sont très nombreuses) que du point de vue explicatif (interprétation du phénomène, notamment par des considérations biochimiques ou alimentaires), il est du même coup justifié, aux yeux du Biologiste, que nous remplaçons l'hypothèse 3 ci-dessus par la suivante :

Hypothèse 3' : "il existe une relation fonctionnelle entre le logarithme du caractère quantitatif x , et les logarithmes des facteurs $z_1, z_2, \dots, z_k$. Cette relation est linéaire".

En fait, bien que le modèle soit ici du type :

$$\left\{ \begin{array}{l} x_i = K_i \prod_{j=1}^k z_j^{\alpha_{i,j}} \\ i = 1, 2, \dots, n \end{array} \right.$$

nous pourrons toujours effectuer une analyse factorielle linéaire pour trouver les coefficients K_i et $\alpha_{i,j}$, puisque ce modèle est équivalent à celui-ci :

$$\left\{ \begin{array}{l} \log x_i = \log K_i + \sum_{j=1}^k \alpha_{ij} \log z_{ij} \\ i = 1, 2, \dots, n \end{array} \right.$$

Cette fois, les "caractères" considérés seront les logarithmes des mesures effectuées sur les organismes observés (à moins, bien entendu, que l'instrument de mesure soit lui-même gradué en échelle logarithmique) . Les "facteurs" déterminés par le calcul seront en réalité les logarithmes des facteurs cherchés.

Une telle analyse factorielle sera appelée une analyse factorielle log-linéaire. Bien entendu, le choix de la base des logarithmes est sans influence sur les résultats, les logarithmes ne servant ici que d'intermédiaires de calcul. Par conséquent, la base des logarithmes utilisés sera choisie arbitrairement lors des calculs numériques.

TROISIEME PARTIE

EXEMPLES NUMERIQUES TIRES DE LA BIOLOGIE

CHAPITRE I : ANALYSE D'UN ENSEMBLE DE CARACTERES CHEZ UN PUCERON

A - INTRODUCTION

Ce premier exemple numérique d' application de notre méthode d'analyse factorielle à la Biologie a été choisi sans autre but que d'illustrer la théorie qui précède.

Comme nous l'avons souligné plus haut (Deuxième Partie) une analyse factorielle pratiquée sur n' importe quel lot de caractères d'un individu, peut se justifier a priori si l'on émet l'hypothèse que tous ces caractères dépendent de "facteurs" en relation avec la structure génique et les conditions de milieu.

En fait, notre choix des caractères à analyser ne s' est pas effectué ici complètement au hasard, comme cela était théoriquement permis. Comme on le verra, nous avons choisi, sur un Puceron, sept caractères qui sont tous des segments d'appendices. Plusieurs raisons ont présidé à ce choix. D'une part, nous avons cherché des caractères présentant entre eux une unité, une parenté sinon prouvée, du moins supposée. D' autre part, les segments d'appendices ont été choisis, de préférence à d'autres organes, à cause de la facilité de mesure et de la stabilité de définition (d'autres

organes, par exemple, ont des mensurations trop variables au cours du temps, suivant les circonstances). Enfin, ces mêmes segments d'appendices ont fait l'objet d'une étude biométrique, orientée vers d'autres buts, menée par Monsieur L. DE REGGI, qui a élevé et préparé les Pucerons observés (voir DE REGGI, 1963) ; c'est pourquoi la présente analyse factorielle de caractères du Puceron pourrait éventuellement être intégrée dans une étude plus vaste, faisant appel à différentes techniques bio-statistiques.

Monsieur DE REGGI a mis son matériel et sa propre expérience à notre disposition ; qu'il nous permette de lui exprimer ici nos remerciements.

B - LA POPULATION ETUDIEE. LES MESURES

Nous nous sommes intéressés à un Puceron, Myzus persicae (Sulzer). La population étudiée est formée de femelles parthénogénétiques, aptères, adultes, élevées dans les conditions suivantes (voir DE REGGI, 1963) :

- nourriture : petit chou
- température : $20^{\circ} \pm 1^{\circ}\text{C}$
- humidité relative : 80 %
- éclairage artificiel 14 heures par jour (10 heures d'obscurité)

Sur chaque individu d'un échantillon de taille $N = 100$ extrait de cette population, nous avons mesuré $n = 7$ caractères, qui sont des longueurs de segments d'appendices, tous d'un même côté du corps de l'animal. (voir figure 5a). Plus précisément :

X_1, X_2, X_3, X_4 , les quatre segments terminaux de l'antenne ;

X_5, X_6 , les deux segments terminaux de la patte moyenne ;

X_7 , le cornicule.

(voir figure 5b)

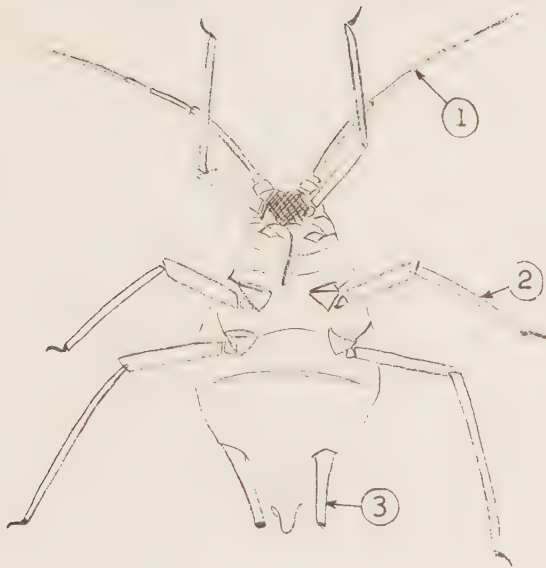
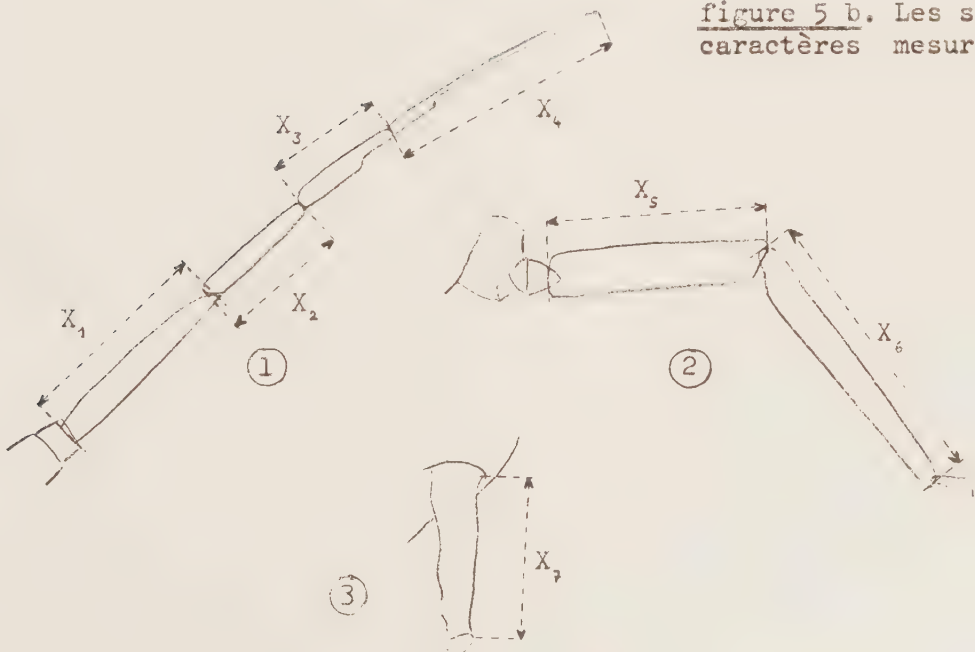


figure 5 a

Myzus persicae, dessin
d'ensemble (d'après
photo L. De Reggi).

figure 5 b. Les sept
caractères mesurés.



- TABLEAU I -

Mesures sur les Pucerons, en unités micrométriques

Réf.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇		X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
I	427	300	222	667	488	857	500	26	441	301	220	675	450	744	474
2	489	358	276	709	540	966	552	27	442	300	245	705	475	785	478
3	447	317	259	637	500	880	541	28	421	304	231	691	458	810	476
4	420	303	228	631	480	843	515	29	481	329	250	670	510	867	541
5	418	300	214	644	488	858	495	30	476	322	229	647	512	867	532
6	487	350	277	709	542	936	538	31	490	318	232	650	527	922	560
7	478	349	252	700	530	941	549	32	477	340	242	660	515	905	558
8	465	308	200	663	487	854	479	33	450	295	230	641	455	817	473
9	476	333	241	700	513	900	543	34	440	313	232	667	500	870	505
10	500	328	260	684	504	850	540	35	421	303	220	671	461	810	478
11	500	323	275	662	527	904	540	36	472	330	236	653	515	906	513
12	488	346	263	710	510	890	515	37	410	300	232	652	458	800	454
13	514	356	280	694	520	925	530	38	417	300	220	696	483	784	464
14	500	357	257	682	504	890	476	39	473	306	246	700	478	820	491
15	500	365	277	698	509	886	500	40	382	273	215	681	447	747	436
16	484	325	263	681	500	879	469	41	420	290	231	669	470	792	459
17	400	286	223	703	416	763	427	42	483	331	239	690	546	892	547
18	420	311	225	689	500	805	500	43	430	277	208	627	458	782	467
19	370	257	214	683	400	703	382	44	489	293	218	586	474	840	478
20	500	355	265	700	547	939	547	45	435	300	215	663	456	818	468
21	391	290	223	700	442	753	443	46	447	318	255	731	496	859	517
22	367	247	220	670	395	684	400	47	455	285	247	692	476	837	482
23	500	351	285	682	520	898	560	48	413	310	243	693	452	800	465
24	489	335	263	700	523	907	550	49	454	339	258	698	494	880	516
25	466	318	247	667	500	867	557	50	451	313	239	689	500	865	484

- TABLEAU I (suite) -

Mesures sur les Pucerons, en unités micrométriques

Réf.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	Réf.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
51	451	324	219	682	512	900	495	76	418	295	232	675	441	772	461
52	443	300	219	690	477	806	462	77	381	250	207	652	390	692	405
53	457	315	238	684	500	876	474	78	442	293	238	688	425	773	444
54	452	325	241	708	500	876	488	79	360	249	187	620	402	685	455
55	459	338	250	700	511	871	521	80	388	272	210	659	421	732	445
56	361	242	208	622	415	735	467	81	374	226	200	590	405	700	429
57	400	279	213	654	417	716	442	82	380	251	200	654	407	703	441
58	400	290	221	674	458	758	439	83	412	263	216	653	421	731	446
59	443	292	249	710	480	847	482	84	389	209	219	669	405	720	416
60	431	300	234	688	460	809	454	85	400	270	214	677	415	740	418
61	500	352	254	756	526	917	516	86	400	275	237	654	423	748	443
62	461	300	245	678	475	828	475	87	393	266	222	667	413	735	431
63	471	323	258	712	516	872	534	88	400	279	226	633	442	804	459
64	400	300	221	688	410	742	434	89	450	308	235	667	468	838	517
65	450	315	241	717	457	798	481	90	415	300	245	669	444	779	476
66	475	320	274	704	495	887	491	91	370	249	200	610	411	700	424
67	446	332	261	738	452	809	441	92	389	263	210	652	413	729	417
68	415	291	237	690	437	770	430	93	400	290	228	686	422	743	396
69	465	324	254	686	503	876	533	94	493	301	265	689	526	919	544
70	483	309	253	662	534	900	568	95	446	388	245	707	476	800	462
71	500	275	276	685	530	931	574	96	453	323	241	716	489	849	527
72	465	291	228	676	463	773	473	97	507	364	267	700	530	924	560
73	371	259	211	661	405	700	441	98	546	368	298	729	560	968	581
74	429	300	243	687	454	774	454	99	535	355	274	730	556	967	568
75	440	291	239	664	460	805	491	100	536	366	281	646	551	961	574

C - ANALYSE FACTORIELLE LINEAIRE

Les moyennes, variances et covariances de l'échantillon considéré ont été calculées, à partir des données ci-dessus, au moyen d'une "Mercator" (Olivetti) programmée à cet effet (Laboratoire de Zoologie Expérimentale, Faculté des Sciences de Lyon). On a obtenu (*) :

$$t_{\mu} = (\quad 444,11 \quad 306,18 \quad 238,26 \quad 677,70 \quad 475,54 \quad 826,85 \quad 487,63)$$

$$S = \begin{pmatrix} 1859,38 & 1198,13 & 804,48 & 524,13 & 1711,86 & 2949,15 & 1714,74 \\ & 1119,67 & 579,73 & 544,41 & 1220,61 & 2086,68 & 1139,29 \\ & & 509,33 & 381,21 & 752,73 & 1338,01 & 759,36 \\ & & & 863,53 & 519,86 & 870,31 & 350,64 \\ & & & & 1909,87 & 3161,70 & 1885,10 \\ & & & & & 5605,11 & 3163,82 \\ & & & & & & 2231,41 \end{pmatrix}$$

D'où : $s_E^2 = 14098,30$, et

$$T_G = \begin{pmatrix} 12238,92 & -1198,13 & -804,48 & -524,13 & -1711,86 & -2949,15 & -1714,74 \\ & 12978,63 & -579,73 & -544,41 & -1220,61 & -2086,68 & -1139,29 \\ & & 13588,97 & -381,21 & -752,73 & -1338,01 & -759,36 \\ & & & 13234,77 & -519,86 & -870,31 & -350,64 \\ & & & & 12188,43 & -3161,70 & -1885,10 \\ & & & & & 8493,19 & -3163,82 \\ & & & & & & 11866,89 \end{pmatrix}$$

(*) Dans ce qui suit, lorsqu'une matrice carrée sera écrite d'une manière incomplète, il est sous-entendu qu'il s'agira d'une matrice symétrique (matrices S et matrices T_G).

La détermination des valeurs propres et des vecteurs propres unitaires de la matrice T_G a été effectuée par un ordinateur "I620 IBM" (Centre de Calcul Numérique, Faculté des Sciences de Lyon). Les valeurs propres sont :

$$\left\{ \begin{array}{l} \lambda_1 = 1863,63 \\ \lambda_2 = 13186,04 \\ \lambda_3 = 13798,40 \\ \lambda_4 = 13837,39 \\ \lambda_5 = 13889,89 \\ \lambda_6 = 13978,84 \\ \lambda_7 = 14035,61 \end{array} \right. \quad d' où \quad \left\{ \begin{array}{l} \lambda_1' = 12234,67 \\ \lambda_2' = 912,26 \\ \lambda_3' = 299,90 \\ \lambda_4' = 260,91 \\ \lambda_5' = 208,41 \\ \lambda_6' = 119,46 \\ \lambda_7' = 62,69 \end{array} \right.$$

Les vecteurs propres correspondants $v_1, v_2, \dots, v_7$ forment la matrice C (matrice de changement de base) :

$$C = \begin{pmatrix} 0,36885 & 0,06551 & -0,32178 & -0,66190 & 0,37012 & 0,38349 & -0,18485 \\ 0,26278 & 0,32513 & -0,27183 & -0,21406 & -0,82353 & -0,14712 & -0,07534 \\ 0,16881 & 0,21276 & 0,04281 & -0,29728 & 0,29158 & -0,73892 & 0,45278 \\ 0,11630 & 0,82561 & 0,43991 & 0,17037 & 0,15366 & 0,18902 & -0,15152 \\ 0,38722 & -0,06125 & 0,07023 & 0,15228 & -0,11425 & 0,44573 & 0,77875 \\ 0,66895 & -0,09723 & -0,27904 & 0,55684 & 0,18541 & -0,21125 & -0,27588 \\ 0,39454 & -0,38717 & 0,73785 & -0,25575 & -0,16952 & -0,07910 & -0,22276 \end{pmatrix}$$

Nous avons obtenu tous les éléments nous permettant de calculer l'importance relative de chacun des sept facteurs $Z_1, Z_2, \dots, Z_7$, pour chacune des variables $X_1, X_2, \dots, X_7$. Ces importances relatives, du type

$\frac{a_{kj}^2 \lambda_k'}{s_j^2}$ ont été calculées à l'aide d'une machine de bureau ordinaire. Les

résultats, exprimés en pourcentages, sont consignés dans le tableau II ci-après.

Remarque : Bien entendu, la précision de ces résultats (deux décimales pour un pourcentage) est illusoire ; l'état actuel des techniques de l'analyse factorielle ne nous permet pas d'apprécier valable-

ment des variations de pourcentages de l'ordre de l'unité, et a fortiori de l'ordre du centième. Il nous suffit ici de connaître des ordres de grandeur afin de décider du rejet de certains facteurs jugés sans importance par rapport à d'autres. Nous tenons néanmoins à conserver une telle "précision" tant que nous avons à manipuler ces pourcentages, ceci afin de ne pas introduire d'erreur supplémentaire. Ce n'est qu'à la fin de chaque étude que nous donnerons des résultats arrondis.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
X_1	89,38	0,21	1,67	6,15	1,54	0,94	0,12
X_2	75,45	8,61	1,98	1,07	12,62	0,23	0,03
X_3	68,45	8,11	0,11	4,53	3,48	12,81	2,52
X_4	19,16	72,01	6,72	0,88	0,57	0,49	0,17
X_5	96,05	0,18	0,08	0,32	0,14	1,24	1,99
X_6	97,68	0,15	0,42	1,44	0,13	0,10	0,09
X_7	85,35	6,13	7,32	0,76	0,27	0,03	0,14

- TABLEAU II -

Revenons à l'analyse elle-même. Le tableau des pourcentages nous permet de conserver a priori les facteurs suivants :

- X_1 : Z_1 et Z_4 (95,53 %)
- X_2 : Z_1, Z_2, Z_5 (96,68 %)
- X_3 : Z_1, Z_2, Z_4, Z_5, Z_6 (97,38 %)
- X_4 : Z_1, Z_2, Z_3 (97,89 %)
- X_5 : Z_1 (96,05 %)
- X_6 : Z_1 (97,68 %)
- X_7 : Z_1, Z_2, Z_3 (98,80 %)

Discussion : Le facteur Z_1 est un facteur commun aux sept variables (facteur général), généralement le plus important pour chacune de ces variables (sauf pour X_4). On remarque en particulier que Z_1 est le seul facteur dont dépendent X_5 et X_6 (segments d'une patte). Les quatre variables X_1 à X_4 (segments d'une antenne) dépendent chacune de plusieurs facteurs :

- Z_1 , très important pour X_1 , a de moins en moins d'importance pour X_2 , X_3 , X_4 : son importance va en décroissant au fur et à mesure que l'on s'éloigne du corps de l'animal ;

- Z_2 , par contre, présente le phénomène inverse : d'importance nulle pour X_1 , il apparaît nettement dans X_2 et X_3 , pour devenir le plus important dans X_4 ;

- Z_3 , Z_4 , Z_5 , Z_6 , sont des facteurs plus ou moins spécifiques des quatre variables X_1 à X_4 , apparaissant d'une manière désordonnée rendant impossible leur interprétation dans la présente étude.

- La variable X_4 enfin (cornicule) dépend, bien entendu, et pour une large part, du facteur général Z_1 , et aussi des facteurs Z_2 (dont nous avons vu le rôle qu'il joue sur les segments d'antenne) et Z_3 (facteur dont dépend aussi X_4).

Conclusion : Le modèle retenu à la suite de cette analyse factorielle linéaire est le suivant :

$$\left\{ \begin{array}{l} X_1 = \bar{X}_1 + 0,4 Z_1 - 0,7 Z_4 \\ X_2 = \bar{X}_2 + 0,3 Z_1 + 0,3 Z_2 - 0,8 Z_5 \\ X_3 = \bar{X}_3 + 0,2 Z_1 + 0,2 Z_2 - 0,3 Z_4 + 0,3 Z_5 - 0,7 Z_6 \\ X_4 = \bar{X}_4 + 0,1 Z_1 + 0,8 Z_2 + 0,4 Z_3 \\ X_5 = \bar{X}_5 + 0,4 Z_1 \\ X_6 = \bar{X}_6 + 0,7 Z_1 \\ X_7 = \bar{X}_7 + 0,4 Z_1 - 0,4 Z_2 + 0,7 Z_3 \end{array} \right.$$

Remarque : cet exemple d'analyse factorielle nous permet d'illustrer ce que nous disions dans la Deuxième Partie (Chapitre 2, § D), au sujet de la façon de définir l' "importance" d'un facteur. Si,

nous en tenant à la méthode classique, nous avons considéré seulement les importances globales (les λ_i), nous aurions retenu les deux ou trois premiers facteurs seulement, et nous aurions exprimé chacune des sept variables comme fonction linéaire des deux ou trois premiers facteurs, en utilisant pour cela les trois premières colonnes de la matrice C. Ce faisant, il ne nous serait pas apparu nettement que, par exemple pour X_5 et X_6 , seul le facteur Z_1 a de l'importance ; et inversement, pour X_3 , le facteur Z_5 (négligeable du point de vue de l'importance globale) est bien plus important que Z_3 (non négligé globalement, mais négligeable en fait) . C'est d'ailleurs là un signe de la spécificité d'un facteur : faible importance du point de vue global, mais en réalité cette importance est concentrée sur une seule variable.

D - ANALYSE FACTORIELLE LOG-LINEAIRE

Nous avons effectué les mêmes calculs que précédemment, cette fois sur les logarithmes décimaux des mesures. Le vecteur moyen μ , la matrice des covariances S, sont maintenant :

$${}^t\mu = \begin{bmatrix} 2,64540 & 2,48329 & 2,37511 & 2,83064 & 2,67534 & 2,91567 & 2,68607 \end{bmatrix}$$

$$S = 10^{-6} \begin{bmatrix} 1808,74 & 1727,07 & 1435,36 & 339,74 & 1565,62 & 1548,44 & 1508,23 \\ & 2375,11 & 1526,46 & 515,00 & 1655,29 & 1624,47 & 1495,42 \\ & & 1685,72 & 457,85 & 1263,03 & 1290,78 & 1219,20 \\ & & & 365,47 & 312,71 & 302,16 & 203,28 \\ & & & & 1637,10 & 1554,45 & 1558,68 \\ & & & & & 1582,06 & 1501,16 \\ & & & & & & 1783,11 \end{bmatrix}$$

D'où

$$s_E^2 = 10^{-6} \times 11237,31$$

et :

$$T_G = 10^{-6} \begin{pmatrix} 9428,57 & -1727,07 & -1435,36 & -339,74 & -1565,62 & -1548,44 & -1508,23 \\ 8862,20 & -1526,46 & -515,00 & -1655,29 & -1624,47 & -1495,42 & \\ 9551,59 & -457,85 & -1263,03 & -1290,78 & -1219,20 & & \\ 10871,84 & -312,71 & -302,16 & -203,28 & & & \\ 9600,21 & -1554,45 & -1558,68 & & & & \\ 9655,25 & -1501,16 & & & & & \\ 9454,20 & & & & & & \end{pmatrix}$$

La détermination des valeurs propres et des vecteurs propres unitaires de la matrice T_G , a donné les résultats suivants :

$$\left\{ \begin{array}{l} \lambda_1 = 1\ 808,76 \times 10^{-6} \\ \lambda_2 = 10\ 471,24 \times 10^{-6} \\ \lambda_3 = 10\ 757,61 \times 10^{-6} \\ \lambda_4 = 11\ 020,67 \times 10^{-6} \\ \lambda_5 = 11\ 059,40 \times 10^{-6} \\ \lambda_6 = 11\ 116,05 \times 10^{-6} \\ \lambda_7 = 11\ 190,13 \times 10^{-6} \end{array} \right. \quad \text{d'où} \quad \left\{ \begin{array}{l} \lambda_1' = 9\ 428,55 \times 10^{-6} \\ \lambda_2' = 766,07 \times 10^{-6} \\ \lambda_3' = 479,70 \times 10^{-6} \\ \lambda_4' = 216,64 \times 10^{-6} \\ \lambda_5' = 177,91 \times 10^{-6} \\ \lambda_6' = 121,26 \times 10^{-6} \\ \lambda_7' = 47,18 \times 10^{-6} \end{array} \right.$$

$$C = \begin{pmatrix} 0,41780 & -0,04561 & 0,12730 & -0,76442 & 0,02461 & 0,46960 & 0,04118 \\ 0,45744 & 0,46149 & -0,67111 & 0,09622 & -0,34346 & -0,00663 & 0,01166 \\ 0,36671 & 0,43841 & 0,71554 & 0,14204 & -0,28052 & -0,22160 & -0,11557 \\ 0,09606 & 0,43326 & 0,05050 & 0,32746 & 0,69991 & 0,42640 & 0,14694 \\ 0,40197 & -0,24944 & -0,10286 & 0,04041 & 0,39202 & -0,24365 & -0,74225 \\ 0,39593 & -0,19578 & -0,01803 & -0,07367 & 0,31087 & -0,55749 & 0,62588 \\ 0,39262 & -0,55181 & 0,08925 & 0,52149 & -0,25010 & 0,42231 & 0,14337 \end{pmatrix}$$

A partir de ces résultats, nous pouvons dresser le tableau des importances relatives des n facteurs pour chacune des n variables. Mais auparavant, convenons de certaines notations. Comme précédemment, X_j désigne la mesure du j -ième caractère étudié (mesure en échelle métrique) ; on posera $X_j' = \log X_j$; les calculs ci-dessus ont donc été effectués à partir de la matrice des covariances des X_j' . De plus, les facteurs que nous allons obtenir maintenant n'ont aucun rapport avec les facteurs Z_k déterminés dans l'analyse factorielle linéaire précédente ; pour éviter toute confusion, nous noterons Y_k les facteurs qui nous intéressent maintenant : le modèle obtenu sera donc du type :

$$X_j = K_j Y_1^{\alpha_1} Y_2^{\alpha_2} \dots Y_n^{\alpha_n}$$

$$\text{ou } \log X_j = \log K_j + \alpha_1 \log Y_1 + \dots + \alpha_n \log Y_n ;$$

$$\text{notons } Y_k' = \log Y_k, \quad \text{d'où :}$$

$$X_j' = \log K_j + \alpha_1 Y_1' + \dots + \alpha_n Y_n'.$$

Le présent type d'analyse factorielle nous permet donc de déterminer les "facteurs" Y_k' , les coefficients α_k , et les constantes $\log K_j$ (on a $\log K_j = \overline{X_j'}$).

Les importances relatives des Y_k' pour les variables X_j' sont données par le tableau III.

	Y_1'	Y_2'	Y_3'	Y_4'	Y_5'	Y_6'	Y_7'
X_1'	90,99	0,09	0,43	7,00	0,01	1,48	0,00
X_2'	83,07	6,87	9,10	0,08	0,88	0,00	0,00
X_3'	75,21	8,73	14,57	0,26	0,83	0,35	0,04
X_4'	23,80	39,35	0,33	6,36	23,85	6,03	0,28
X_5'	93,06	2,91	0,31	0,02	1,67	0,44	1,59
X_6'	93,42	1,86	0,01	0,07	1,09	2,38	1,17
X_7'	81,51	13,08	0,21	3,30	0,62	1,21	0,05

Discussion : Nous retiendrons les facteurs suivants :

X_1'	: Y_1' et Y_4'	(97,99 %)
X_2'	: Y_1' , Y_2' et Y_3'	(99,04 %)
X_3'	: Y_1' , Y_2' et Y_3'	(98,51 %)
X_4'	: Y_1' , Y_2' , Y_4' , Y_5' et Y_6'	(99,37 %)
X_5'	: Y_1'	(93,06 %)
X_6'	: Y_1'	(93,42 %)
X_7'	: Y_1' et Y_2'	(94,59 %)

Nous observons à peu près les mêmes phénomènes que dans l'analyse factorielle linéaire précédente : Y_1' est un facteur commun aux sept variables, généralement le plus important pour chacune de ces variables (sauf pour X_4'). Ce facteur Y_1' est pratiquement le seul ayant une action sur X_5' et X_6' (segments de la patte). En ce qui concerne les quatre segments d'antenne (X_1' à X_4'), Y_1' , toujours très important, voit cependant son rôle diminuer au fur et à mesure que l'on s'éloigne du corps de l'animal ; le facteur Y_2' possède une propriété analogue, mais en sens inverse : son importance, nulle pour X_1' , s'accroît en même temps que l'éloignement du corps, et atteint son maximum en X_4' , à l'extrémité de l'antenne.

Les facteurs Y_3' , Y_4' , Y_5' et Y_6' jouent des rôles plus ou moins importants pour les variables X_1' à X_4' . Là encore, leur apparition se fait d'une manière désordonnée, ce qui rend impossible leur interprétation d'après ces seuls résultats.

Conclusion : Le modèle résultant de l'analyse ci-dessus est le suivant :

$$\begin{cases}
 X_1' = \overline{X_1'} + 0,4 Y_1' & - 0,8 Y_4' \\
 X_2' = \overline{X_2'} + 0,5 Y_1' + 0,5 Y_2' - 0,7 Y_3' \\
 X_3' = \overline{X_3'} + 0,4 Y_1' + 0,4 Y_2' + 0,7 Y_3' \\
 X_4' = \overline{X_4'} + 0,1 Y_1' + 0,4 Y_2' & + 0,3 Y_4' + 0,7 Y_5' + 0,4 Y_6' \\
 X_5' = \overline{X_5'} + 0,4 Y_1' \\
 X_6' = \overline{X_6'} + 0,4 Y_1' \\
 X_7' = \overline{X_7'} + 0,4 Y_1' - 0,6 Y_2'
 \end{cases}$$

soit encore :

$$\left\{ \begin{array}{l} X_1 = 442 Y_1^{0,4} Y_4^{-0,8} \\ X_2 = 305 Y_1^{0,5} Y_2^{0,5} Y_3^{-0,7} \\ X_3 = 237 Y_1^{0,4} Y_2^{0,4} Y_3^{0,7} \\ X_4 = 677 Y_1^{0,1} Y_2^{0,4} Y_4^{0,3} Y_5^{0,7} Y_6^{0,4} \\ X_5 = 474 Y_1^{0,4} \\ X_6 = 823 Y_1^{0,4} \\ X_7 = 485 Y_1^{0,4} Y_2^{-0,6} \end{array} \right.$$

E - CONCLUSIONS :

Nous avons pratiqué, sur un même ensemble de données numériques, deux types différents d'analyse factorielle : linéaire et log-linéaire.

Dans les deux cas, nous avons dégagé les faits essentiels suivants :

- il existe un facteur général, commun aux sept variables étudiées, rendant compte d'une part très importante de la variation de celles-ci (facteur noté Z_1 dans un cas, Y_1 dans l'autre cas).

- ce facteur est pratiquement le seul ayant une action sur les segments de la patte (X_5 et X_6).

- ce facteur a une action de moins en moins grande sur les segments de l'antenne (X_1 à X_4) au fur et à mesure que l'on s'éloigne du corps.

- parallèlement, il existe un second facteur (Z_2 ou Y_2) sans importance marquée pour la patte, mais important pour l'antenne et le cornicule ; ce facteur a une action de plus en plus grande sur les segments de l'antenne au fur et à mesure que l'on s'éloigne du corps. Dans les deux cas, ce facteur s'avère le plus important pour le dernier segment de l'antenne (X_4).

Si, intuitivement, on admet que les "extrémités" sont, plus que

le reste du corps, sensibles aux conditions de milieu, cela nous conduit à penser que le facteur Z_2 doit dépendre du milieu, alors que Z_1 doit être d'ordre purement génique, soit directement, soit par un intermédiaire physiologique (par exemple, la présence d'une hormone, dont la concentration irait en décroissant en fonction de l'éloignement du centre du corps). Le fait que, contrairement à la patte, seuls le cornicule et l'antenne sont sensibles au facteur "milieu", montre peut-être une certaine faculté d'adaptation de ces deux organes (finalité de la forme ?), faculté que n'a pas (ou "dont n'a pas besoin") la patte.

Quant aux autres facteurs, il ne nous est pas possible de les interpréter ici.

Peut-être des mesures faites sur des Pucerons élevés dans différents milieux, et aussi au cours de leur croissance, permettraient-elles de résoudre (ou simplement de mieux orienter) les quelques questions biologiques suggérées ci-dessus par le modèle mathématique obtenu à la suite de l'analyse factorielle.

C H A P I T R E I I : LES FACTEURS DE LA FORME DE LA FEUILLE DE VIGNE

A - INTRODUCTION

Très souvent, un être vivant peut se reconnaître dès le premier abord au moyen d'un très petit nombre de caractères (forme, comportement..). Parmi ces caractères, l'un des plus importants est sans doute la forme. En effet, ce concept de forme permet de reconnaître d'un simple coup d'oeil les animaux et plantes communs (animaux domestiques, fleurs, arbres). C'est par des différences de forme qu'on distingue des espèces voisines.

Le Naturaliste, poussant plus loin cette connaissance première de l'Etre par sa forme, en examine tous les détails et opère des classifications de plus en plus raffinées. Mais en fin de compte, l'observateur se réfère le plus souvent, pour établir de telles classifications, à des dé-

tails visibles sur l'organisme (que ce soit à l'oeil nu ou au microscope, avec ou sans dissection), donc surtout à des détails de forme.

Nous n'insisterons pas davantage sur cette question de Philosophie de la Connaissance en général, et de Méthodologie des Sciences Naturelles en particulier. L'essentiel est de reconnaître que la forme d'un être est un élément fondamental de cet être (tout au moins, à l'échelle des Métazoaires).

Le caractère "forme" peut se définir comme un ensemble de caractères quantitatifs (longueurs), en nombre d'autant plus élevé que cette forme est plus compliquée. Si tous ces caractères peuvent eux-mêmes être exprimés en fonction d'un petit nombre de facteurs, alors ces quelques facteurs définissent aussi la forme de l'être ; et comme la forme est un élément fondamental pour l'être, nous pourrions penser que ces facteurs sont également des facteurs fondamentaux pour l'être, globalement.

Cette idée d'importance de la forme, d'une part, de résolution de la forme en un certain nombre de caractères quantitatifs, d'autre part, nous a amenés à essayer d'appliquer l'analyse factorielle à l'étude de la forme d'un organisme. Nous avons choisi ici comme exemple un cas où l'importance de la forme (pour la caractérisation de l'organisme) est particulièrement nette, puisqu'il s'agit de la feuille d'un végétal.

B - LA POPULATION ETUDIEE. LES MESURES

Nous avons prélevé au hasard cent feuilles (intactes) sur un même plant de Vigne, cultivé aux environs de Montpellier (Vitis vinifera).

Nous avons constaté que ces feuilles se ressemblaient en ce sens qu'elles avaient toutes à peu près la même forme. Comme l'ont montré les mesures, elles ne sont pas rigoureusement semblables entre elles (au sens de la similitude géométrique) ; mais on peut penser que la forme de chaque feuille provient d'une petite "déformation" de la forme idéale de la feuille de Vigne - forme que nous ne connaissons pas, mais dont nous pouvons choisir une approximation supposée bonne, en prenant par exemple la forme moyenne des cent feuilles considérées.

Une caractérisation complète de la forme de la feuille de Vigne nécessiterait de nombreuses mesures. Mais il ne s'agit présentement que de

donner un exemple d'application d' une méthode d'analyse factorielle. C'est pourquoi nous nous sommes limités ici à dix mesures par feuille, mesures qui nous semblent représenter assez bien la forme de la feuille (sans détails toutefois) (voir figure 6a).

Les mesures ont été faites au moyen d' une réglette graduée en millimètres. La face antérieure de la feuille étant tournée vers l'observateur, nous avons mesuré : le pétiole (X_1), les cinq grosses nervures (X_2 , X_3 , X_4 , X_5 , X_6), quatre distances entre les extrémités de ces cinq nervures (X_7 , X_8 , X_9 , X_{10}) (voir figure 6b).



figure 6

- a) une feuille de Vigne (reproduction schématisée en vraie grandeur)
b) schéma montrant la disposition des dix caractères mesurés.

- TABLEAU I -

Mesures sur les feuilles de Vigne, en millimètres

Réf	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	Réf	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀
1	35	120	100	100	72	66	92	87	95	57	26	65	88	77	82	55	55	60	68	65	60
2	38	59	50	53	34	37	41	44	45	46	27	30	40	40	36	28	28	38	26	36	22
3	94	112	100	100	73	67	87	77	55	97	28	28	42	35	35	22	20	33	33	30	25
4	37	58	48	50	34	35	44	33	47	54	29	38	48	45	37	28	28	37	22	35	40
5	34	55	45	49	36	31	43	43	40	34	30	55	87	80	74	58	53	60	52	60	56
6	50	64	61	66	48	49	36	48	58	51	31	96	115	110	105	90	83	105	93	110	74
7	50	74	61	69	47	47	35	52	62	62	32	60	82	80	70	57	54	59	48	61	55
8	86	108	99	91	77	67	93	93	72	67	33	56	80	66	67	50	52	50	60	50	60
9	48	64	60	60	47	47	50	55	51	51	34	60	81	72	78	50	55	60	55	63	60
10	32	51	50	49	35	33	45	35	31	35	35	52	72	72	62	55	47	56	46	55	45
11	40	62	53	56	42	47	36	54	45	44	36	48	73	65	70	53	55	48	49	37	60
12	65	68	65	63	51	45	52	47	56	52	37	52	75	65	73	50	52	43	59	57	55
13	68	80	70	70	52	56	63	59	51	69	38	55	70	65	58	55	46	54	38	52	45
14	72	72	65	66	49	52	46	53	68	50	39	19	35	31	32	22	19	31	29	21	27
15	37	40	40	39	25	24	34	33	31	34	40	25	43	40	42	24	29	36	40	30	34
16	48	60	52	50	40	43	41	52	49	43	41	26	47	40	42	32	29	35	35	35	30
17	22	33	32	32	22	22	24	24	27	31	42	47	66	65	65	50	46	58	57	34	46
18	46	51	48	48	39	38	45	25	42	38	43	3	61	56	50	45	37	38	33	42	41
19	31	60	52	52	34	37	44	41	47	47	44	50	63	53	55	41	42	45	48	52	52
20	15	33	31	27	24	24	31	22	25	20	45	65	70	65	65	47	48	45	45	54	57
21	18	33	31	31	17	20	31	26	29	25	46	35	49	42	42	30	30	38	36	34	33
22	57	72	67	74	46	51	48	57	54	45	47	19	34	30	29	21	20	21	25	21	23
23	52	65	62	57	52	48	37	39	57	50	48	35	45	45	51	32	39	38	42	35	34
24	34	43	36	36	25	27	33	37	34	24	49	59	63	60	60	50	46	55	38	40	59
25	62	73	67	65	52	49	52	56	65	47	50	63	69	65	71	51	54	50	58	61	51

- TABLEAU I (suite) -

Mesures sur les feuilles de Vigne, en millimètres

Réf	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	Réf	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀
51	30	64	52	53	35	35	45	45	40	41	76	23	49	41	44	33	31	29	35	35	35
52	33	58	54	55	42	43	43	41	22	47	77	56	79	77	66	55	47	65	62	58	53
53	31	53	48	50	35	40	38	43	35	35	78	53	79	66	71	47	48	54	64	54	68
54	85	74	68	72	48	54	56	56	52	55	79	36	60	58	56	43	44	52	39	35	39
55	35	50	48	54	32	36	36	48	42	24	80	37	49	47	46	33	34	34	36	40	31
56	65	68	60	66	45	47	53	48	55	56	81	48	53	54	53	40	42	44	42	25	43
57	52	55	49	49	37	38	44	44	39	41	82	49	80	72	70	57	53	55	45	60	63
58	45	52	51	46	37	35	45	36	40	45	83	43	74	68	67	48	46	55	54	60	55
59	55	59	55	57	42	40	44	43	50	40	84	50	77	72	76	52	52	45	57	51	50
60	82	65	66	61	47	45	58	47	50	47	85	40	73	70	73	52	52	53	53	44	65
61	67	60	57	60	39	45	41	49	38	37	86	42	77	70	73	50	49	52	59	60	61
62	20	43	38	45	24	28	27	40	32	32	87	47	66	64	59	41	35	49	38	56	45
63	27	47	42	40	27	27	36	29	36	30	88	75	72	69	68	46	45	61	59	53	38
64	29	60	50	56	31	35	35	43	40	51	89	61	60	59	55	41	43	44	52	56	34
65	35	58	52	52	37	36	42	32	41	48	90	70	71	68	64	52	50	68	62	64	45
66	60	70	60	64	45	48	53	55	60	49	91	100	67	65	69	48	53	50	52	37	60
67	55	70	62	67	50	50	50	57	54	71	92	34	60	60	57	39	37	47	45	47	41
68	53	62	62	55	49	47	40	41	46	48	93	62	73	75	71	54	52	65	55	53	45
69	49	60	57	57	44	42	51	46	24	61	94	43	45	43	40	29	29	36	37	41	33
70	92	76	68	68	51	50	60	60	57	60	95	65	66	67	63	50	49	47	45	62	64
71	85	65	66	60	42	45	58	53	55	40	96	50	75	57	71	47	52	50	52	53	62
72	50	44	40	44	29	28	39	45	34	34	97	55	77	76	77	57	60	52	59	63	52
73	64	63	63	58	48	42	53	42	52	41	98	42	64	66	59	48	41	56	48	52	41
74	88	75	70	70	59	57	57	59	54	58	99	30	64	55	60	41	46	40	42	35	55
75	52	49	40	48	30	32	29	40	35	41	100	41	72	60	64	43	48	48	45	53	56

C - ANALYSE FACTORIELLE LINEAIRE

Les calculs ont été effectués avec les mêmes moyens que dans l'exemple précédent. Nous avons obtenu :

$$t_{\mu} = \left[\begin{array}{cccccccccccc} 49,89 & 63,90 & 58,66 & 58,83 & 43,20 & 42,92 & 47,60 & 46,89 & 47,36 & 46,83 \end{array} \right]$$

$$10^2 S = \left[\begin{array}{cccccccccccc} 35578 & 22937 & 22249 & 21152 & 18040 & 16837 & 19507 & 18347 & 18492 & 15824 \\ 26971 & 23999 & 23514 & 19254 & 17438 & 19859 & 19386 & 19812 & 18054 & \\ 22840 & 21348 & 18204 & 16148 & 18822 & 17357 & 18049 & 15669 & & \\ 21900 & 17112 & 16166 & 17092 & 17830 & 17200 & 16350 & & & \\ 15422 & 13515 & 14942 & 13769 & 14517 & 13079 & & & & \\ 12993 & 12776 & 13017 & 13018 & 12432 & & & & & \\ 18712 & 15133 & 14743 & 12207 & & & & & & \\ 17660 & 14640 & 12171 & & & & & & & \\ & 20483 & 11445 & & & & & & & \\ & & 17442 & & & & & & & \end{array} \right]$$

D'où $s_E^2 = 2100,01$ et

$$10^2 T_G = \left[\begin{array}{cccccccccccc} 174423 & -22937 & -22249 & -21152 & -18040 & -16837 & -19507 & -18347 & -18492 & -15824 \\ 183030 & -23999 & -23514 & -19254 & -17438 & -19859 & -19386 & -19812 & -18054 & \\ 187161 & -21348 & -18204 & -16148 & -18822 & -17357 & -18049 & -15669 & & \\ 188101 & -17112 & -16166 & -17092 & -17830 & -17200 & -16350 & & & \\ 194579 & -13515 & -14942 & -13769 & -14517 & -13079 & & & & \\ 197008 & -12776 & -13017 & -13018 & -12432 & & & & & \\ 191289 & -15133 & -14743 & -12207 & & & & & & \\ 192341 & -14640 & -12171 & & & & & & & \\ & 189518 & -11445 & & & & & & & \\ & & 192559 & & & & & & & \end{array} \right]$$

$$\begin{array}{ccc}
 \left(\begin{array}{l} \lambda_1 = 329,52 \\ \lambda_2 = 1979,61 \\ \lambda_3 = 2022,42 \\ \lambda_4 = 2051,55 \\ \lambda_5 = 2063,55 \\ \lambda_6 = 2076,90 \\ \lambda_7 = 2089,73 \\ \lambda_8 = 2093,73 \\ \lambda_9 = 2095,32 \\ \lambda_{10} = 2097,76 \end{array} \right. & \text{d'où} & \left(\begin{array}{l} \lambda_1' = 1770,49 \\ \lambda_2' = 120,40 \\ \lambda_3' = 77,59 \\ \lambda_4' = 48,46 \\ \lambda_5' = 36,46 \\ \lambda_6' = 23,11 \\ \lambda_7' = 10,28 \\ \lambda_8' = 6,28 \\ \lambda_9' = 4,69 \\ \lambda_{10}' = 2,25 \end{array} \right)
 \end{array}$$

$10^4 C =$

3803	-8987	1515	1024	281	57	-1021	389	-279	255
3804	2529	-42	-199	-250	484	-5784	3112	-5920	-777
3511	1142	-643	-1331	-2587	-2847	-2278	1863	6449	-4375
3415	1853	971	123	2587	-2391	-3022	-5980	1847	4820
2842	955	126	-286	-2796	-3842	4293	4365	-578	5524
2597	723	1151	921	596	-4335	4286	-3542	-3824	-5086
2956	-180	-1932	-6120	-4034	4339	1811	-3267	-957	288
2870	662	-1459	-3025	7809	1769	2527	2811	1141	-596
2926	713	-6282	6414	-662	2707	1265	-753	622	73
2592	2384	7048	2908	-781	4806	1763	444	1522	-371

D'où le tableau II des importances relatives des 10 facteurs pour chacune des 10 variables.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
X_1	71,97	27,33	0,50	0,14	0,01	0,00	0,03	0,00	0,00	0,00
X_2	94,99	2,86	0,00	0,01	0,01	0,02	1,28	0,23	0,61	0,01
X_3	95,56	0,69	0,14	0,38	1,07	0,82	0,23	0,10	0,85	0,19
X_4	95,28	1,89	0,33	0,00	1,11	0,60	0,43	1,03	0,07	0,24
X_5	92,73	0,71	0,01	0,03	1,85	2,21	1,23	0,78	0,01	0,45
X_6	91,90	0,48	0,79	0,32	0,10	3,34	1,45	0,61	0,53	0,45
X_7	82,68	0,02	1,55	9,70	3,17	2,33	0,18	0,36	0,02	0,00
X_8	82,59	0,30	0,94	2,51	12,59	0,41	0,37	0,28	0,03	0,00
X_9	74,00	0,30	14,95	9,73	0,08	0,83	0,08	0,02	0,01	0,00
X_{10}	68,20	3,92	22,10	2,35	0,13	3,06	0,18	0,01	0,06	0,00

- TABLEAU 11 -

Ce tableau nous conduit à conserver a priori les facteurs suivants :

- X_1 : Z_1 et Z_2 (99,30 %)
- X_2 : Z_1 (94,99 %)
- X_3 : Z_1 (95,56 %)
- X_4 : Z_1 (94,28 %)
- X_5 : Z_1 , Z_5 , Z_6 , et Z_7 (98,02 %)
- X_6 : Z_1 , Z_6 , et Z_7 (96,69 %)
- X_7 : Z_1 , Z_3 , Z_4 , Z_5 , Z_6 (99,43 %)
- X_8 : Z_1 , Z_4 , Z_5 (97,69 %)
- X_9 : Z_1 , Z_3 , Z_4 (98,68 %)
- X_{10} : Z_1 , Z_2 , Z_3 , Z_4 , Z_6 (99,63 %)

Discussion :

Le facteur Z_1 est incontestablement un facteur commun aux dix caractères étudiés, facteur très important pour chacun d'eux.

Les autres facteurs retenus ont des importances variables suivant les caractères. Pour pouvoir les retenir définitivement, il nous semble nécessaire d'examiner de plus près l'organisme étudié. Celui-ci pouvant être présumé symétrique par rapport à l'axe pétiole-nervure médiane, les autres caractères peuvent être appariés : X_3 et X_4 , X_5 et X_6 , X_7 et X_8 , X_9 et X_{10} . En se basant sur cette symétrie, tout facteur conditionnant l'un des caractères conditionne aussi l'autre (mis à part bien entendu les facteurs "accidentels" concernant l'individu seulement, et non l'espèce dans son ensemble).

Ceci nous conduit à réexaminer le cas des facteurs qui, ayant été retenus pour l'un des deux caractères, ont été rejetés pour l'autre. L'examen du tableau des importances relatives nous montre que de tels facteurs, lorsqu'ils ont été retenus, ont des importances ne dépassant jamais 4 %. Nous pouvons donc opter pour leur rejet pur et simple. Ainsi nous supprimons Z_5 de X_5 , Z_3 et Z_6 de X_7 , Z_2 et Z_6 de X_{10} .

Que reste-t-il après cet élagage ? Outre le facteur général Z_1 , le facteur Z_2 apparaît comme spécifique de la variable X_1 . Des facteurs Z_6 et Z_4 , spécifiques du couple $X_5 - X_6$, nous ne remarquerons que leur faible importance (de l'ordre de 5 % au total, pour chaque variable), qui justifierait une éventuelle suppression de ces facteurs.

Nous ferons quelques remarques au sujet des facteurs Z_3 , Z_4 , et Z_5 . Le facteur Z_4 est commun aux deux couples de caractères $X_7 - X_8$ et $X_9 - X_{10}$. Un fait curieux est que, pour chacun de ces couples, ce même facteur est plus important du côté gauche (X_7 et X_9) que du côté droit (X_8 et X_{10}) (10 % contre 3 % environ). D'autre part, le couple $X_7 - X_8$ admet un facteur spécifique, Z_5 , qui présente la même particularité, mais en sens inverse (12 % à droite contre 3 % à gauche) ; on remarque de plus que l'importance totale des deux facteurs Z_4 et Z_5 est à peu près la même pour X_7 et pour X_8 (13 % contre 15 %) (voir tableau III).

	Z_4	Z_5	
X_7	9,70	3,17	→ 12,87 %
X_8	2,51	12,59	→ 15,10 %

- TABLEAU III -

Il semble donc que le déséquilibre créé par Z_4 tend à être compensé par le facteur Z_5 (ou réciproquement ?). Cette impression se trouve renforcée lorsque, nous reportant au tableau des coefficients (matrice C), nous constatons que ce facteur Z_5 agit dans des sens opposés sur X_7 et X_8 (coefficient négatif pour X_7 , positif pour X_8).

Pour le couple $X_9 - X_{10}$, le même phénomène apparaît ; cette fois, c'est le facteur Z_3 , spécifique du couple, qui compense le déséquilibre créé par Z_4 (25 % contre 25 %) (voir tableau IV).

	Z_3	Z_4	
X_9	14,95	9,73	→ 24,68 %
X_{10}	22,10	2,35	→ 24,45 %

- TABLEAU IV -

Ce facteur Z_3 agit dans des sens opposés sur X_9 et X_{10} (coefficient négatif pour X_9 , positif pour X_{10}).

D - CONCLUSIONS :

Le modèle que nous retenons en définitive, pour exprimer les dix caractères comme fonctions linéaires des facteurs mis en "évidence" est le suivant :

$$\begin{aligned}
 \left\{ \begin{array}{l}
 X_1 = \bar{X}_1 + 0,40 Z_1 - 0,90 Z_2 \\
 X_2 = \bar{X}_2 + 0,40 Z_1 \\
 X_3 = \bar{X}_3 + 0,35 Z_1 \\
 X_4 = \bar{X}_4 + 0,35 Z_1 \\
 X_5 = \bar{X}_5 + 0,30 Z_1 \quad (- 0,40 Z_6 + 0,40 Z_7) \\
 X_6 = \bar{X}_6 + 0,30 Z_1 \quad (- 0,40 Z_6 + 0,40 Z_7) \\
 X_7 = \bar{X}_7 + 0,30 Z_1 \quad - 0,60 Z_4 - 0,40 Z_5 \\
 X_8 = \bar{X}_8 + 0,30 Z_1 \quad - 0,30 Z_4 + 0,80 Z_5 \\
 X_9 = \bar{X}_9 + 0,30 Z_1 \quad - 0,60 Z_3 + 0,60 Z_4 \\
 X_{10} = \bar{X}_{10} + 0,30 Z_1 \quad + 0,70 Z_3 + 0,30 Z_4
 \end{array} \right.
 \end{aligned}$$

Ces dix caractères dépendent donc tous d'un facteur général Z_1 , très important ; ce facteur est pratiquement le seul dont dépendent les longueurs des cinq grosses nervures (X_2 à X_6) ; le pétiole X_1 , en plus du facteur général, dépend d'un seul facteur spécifique Z_2 . Quant aux distances entre extrémités des nervures (variables X_7 à X_{10}), elles dépendent d'un facteur Z_4 ayant une action dissymétrique (plus forte à gauche qu'à droite de la nervure médiane). Cette action dissymétrique de Z_4 est compensée, pour chaque couple de distances, par un facteur spécifique du couple, à action dissymétrique aussi, mais cette fois plus forte à droite qu'à gauche (facteur Z_3 pour le couple $X_9 - X_{10}$, facteur Z_5 pour le couple $X_7 - X_8$).

Le fait que les cinq grosses nervures ne dépendent que d'un seul et même facteur n'a en soit rien d'étonnant ; ce facteur, qui est important aussi pour les distances entre les extrémités de ces nervures (donc pour les angles qu'elles forment entre elles), régit presque entièrement l'ensemble des mesures définissant la forme de la feuille ; ce facteur Z_1 est véritablement un facteur de forme. De même l'existence d'un facteur Z_2 , spécifique du pétiole, pouvait être soupçonnée : la longueur du pétiole n'est pas une caractéristique de la forme de la feuille ; il est d'expérience commune qu'une grande feuille peut avoir un petit pétiole, et réciproquement.

Mais le phénomène le plus curieux - et inattendu - est l'existence

de facteurs Z_3 , Z_4 , Z_5 , chacun ayant une action dissymétrique sur les deux parties de la feuille, mais arrivant à se compenser mutuellement pour rétablir la symétrie globale de la feuille. Ce phénomène mériterait certainement d'être étudié en détail ; une telle étude sortirait cependant du cadre du présent travail.

CHAPITRE III : LES FACTEURS D'HETEROGENEITE DE L'ESPECE STREPTOCOCCUS FAECALIS

A - INTRODUCTION

Ce troisième exemple d'analyse factorielle appliquée à la Biologie a été choisi pour deux raisons :

Tout d'abord, il a été tiré de la bibliographie : il concerne des variables sur lesquelles a déjà été appliquée une autre méthode d'analyse factorielle, dont les résultats ont été publiés (DEFAYOLLE et COLOBERT, 1962).

De plus, à la différence des deux exemples précédents, les variables considérées ne sont pas des caractères quantitatifs mesurés sur des individus d'une population donnée :

- d'une part, les caractères étudiés sont qualitatifs ; nous verrons par quel moyen on peut les "quantifier" afin de pouvoir appliquer notre méthode d'analyse factorielle ;

- d'autre part, la population n'est pas un ensemble d'individus, mais un ensemble de souches différentes de l'espèce Streptococcus faecalis ; ces souches sont susceptibles de différer entre elles par les caractères étudiés ; chaque souche joue alors le rôle d'un "individu" au sens où nous l'avons entendu dans l'exposé théorique ci-dessus.

Le problème est ici celui de la classification bactérienne, "fai-

sant trop souvent appel à des données intuitives peu accessibles à la critique rationnelle".

Les différences constatées entre les souches conduisent à reconnaître une certaine hétérogénéité de l'espèce Streptococcus faecalis, et par conséquent la nécessité d'une classification plus fine. Dans ces conditions, comment va-t-on caractériser les différents groupes biologiques composant l'espèce Streptococcus faecalis ? Les caractères par lesquels les souches diffèrent entre elles étant trop nombreux (et vraisemblablement liés entre eux), ne peuvent servir pratiquement à une telle classification, qui se trouverait extrêmement diversifiée.

M. DEFAYOLLE et L. COLOBERT ont eu l'idée d'utiliser, pour résoudre ce problème, les ressources de l'analyse factorielle pratiquée sur ces caractères ; la classification en groupes s'effectue alors suivant les "facteurs" et non suivant les caractères. La méthode utilisée par ces auteurs est celle de THURSTONE ; trois facteurs sont alors mis en évidence, et il apparaît que les différentes souches étudiées, classées maintenant selon les valeurs prises par ces facteurs pour chacune d'elles, se répartissent en trois groupes distincts que l'on peut identifier aux sous-espèces S. faecalis proprium, S. faecium et S. durans définies par ailleurs par la taxonomie classique.

Cet essai d'utilisation de l'analyse factorielle, à un problème de classification, nous paraissant concluant, nous avons tenu à lui appliquer la méthode d'analyse exposée ci-avant, afin de préciser le mode de raisonnement qu'il est possible d'utiliser dans des cas semblables. Signalons que nous nous en tiendrons ici uniquement à l'aspect factoriel du problème, sans examiner les conséquences que pourraient avoir les résultats d'une telle analyse.

Nous remercions particulièrement Monsieur DEFAYOLLE et Monsieur COLOBERT d'avoir bien voulu nous communiquer certaines données, absentes de leur publication, et néanmoins fondamentales pour permettre l'application de notre méthode d'analyse factorielle.

B - LA POPULATION ETUDIEE. LES MESURES

Rappelons ce qui a été dit dans l'introduction ci-dessus : la po-

pulation étudiée est l'espèce Streptococcus faecalis, considérée comme un ensemble de souches. Celles-ci ont en commun le fait d'appartenir à la même espèce, mais diffèrent entre elles par certains caractères.

Ont été observées : $N = 173$ souches, du point de vue de $n = 14$ caractères qualitatifs ; pour chaque souche, chacun de ces caractères a été noté soit présent, soit absent. Ces quatorze caractères sont les suivants :

- C_1 = Appartenance au groupe D de LANCEFIELD.
- C_2 = Croissance sur un milieu au tellurite de potassium à I/2500.
- C_3 = Réduction du triphényltétrazolium sur milieu de BARNES.
- C_4 = Fermentation du mannitol.
- C_5 = Fermentation du sorbitol.
- C_6 = Fermentation de l'arabinose.
- C_7 = Fermentation du raffinose.
- C_8 = Activité tyrosine-décarboxylasique.
- C_9 = Réduction du lait au bleu de méthylène.
- C_{10} = Forte réduction du lait au bleu de méthylène.
- C_{11} = Coagulation du lait au bleu de méthylène.
- C_{12} = Forte réduction du lait tournesolé.
- C_{13} = Acidification du lait tournesolé.
- C_{14} = Coagulation du lait tournesolé.

Pour un caractère C_j , les N souches se répartissent en deux catégories :

$$\left\{ \begin{array}{ll} p_j & \text{souches où le caractère } C_j \text{ est présent,} \\ q_j & \text{souches où le caractère } C_j \text{ est absent ;} \end{array} \right.$$

on a donc : $p_j + q_j = N$.

D'autre part, pour deux caractères C_j et C_k , les N souches se répartissent en quatre catégories :

$$\left\{ \begin{array}{ll} a_{jk} & \text{souches où } C_j \text{ et } C_k \text{ sont simultanément présents,} \\ b_{jk} & \text{souches où } C_j \text{ est absent, } C_k \text{ présent,} \\ c_{jk} & \text{souches où } C_j \text{ est présent, } C_k \text{ absent,} \\ d_{jk} & \text{souches où } C_j \text{ et } C_k \text{ sont simultanément absents.} \end{array} \right.$$

		C_j	
		présent	absent
C_k	présent	a_{jk}	b_{jk}
	absent	c_{jk}	d_{jk}

$$\text{On a donc : } \left\{ \begin{array}{l} a_{jk} + b_{jk} + c_{jk} + d_{jk} = N \\ a_{jk} + c_{jk} = p_j \\ b_{jk} + d_{jk} = q_j \\ a_{jk} + b_{jk} = p_k \\ c_{jk} + d_{jk} = q_k \end{array} \right.$$

On remarque en particulier que la connaissance de a_{jk} , p_j et p_k suffit à déterminer complètement les huit nombres a_{jk} , b_{jk} , c_{jk} , d_{jk} , p_j , q_j , p_k , q_k . On a alors, pour chaque caractère C_j , déterminé le nombre p_j de souches où ce caractère est présent ; et d'autre part, pour chaque couple de caractères C_j , C_k , on a déterminé le nombre a_{jk} de souches où les deux caractères sont simultanément présents.

Remarque : le nombre p_j peut aussi être noté a_{jj} , conformément à la définition de a_{jk} ; cela revient à identifier le caractère C_j et le couple de caractères C_j , C_j . Dans ces conditions, on peut dire que :

$$a_{jj} = p_j ; b_{jj} = 0 ; c_{jj} = 0 ; d_{jj} = q_j$$

Le tableau I ci-après est celui des a_{jk} ($j = 1, 2, \dots, 14$; $k = 1, 2, \dots, 14$) ; ce tableau comporte en particulier, sur sa diagonale principale, les a_{jj} , c'est-à-dire les p_j .

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉	C ₁₀	C ₁₁	C ₁₂	C ₁₃	C ₁₄
C ₁	I4I	75	7I	I27	68	57	I9	99	II4	I23	58	I03	I20	80
C ₂		82	64	64	58	I4	7	67	69	74	34	75	73	63
C ₃			73	72	56	I2	5	59	6I	68	30	7I	68	59
C ₄				I48	74	67	24	96	II7	I27	62	III	I29	83
C ₅					75	8	6	57	56	60	29	68	69	52
C ₆						72	I7	37	60	67	34	39	57	27
C ₇							33	I2	2I	24	8	I8	27	7
C ₈								II2	99	I05	43	84	I0I	7I
C ₉									I36	I35	64	95	II7	80
C ₁₀										I47	65	I04	I27	84
C ₁₁											67	50	62	46
C ₁₂												I23	II4	8I
C ₁₃													I48	85
C ₁₄														89

- TABLEAU I -

C - ANALYSE FACTORIELLE LINEAIRE

La méthode d'analyse factorielle que nous avons exposée plus haut concerne des variables aléatoires essentiellement quantitatives. Cette méthode est donc inapplicable directement aux caractères qualitatifs C_i.

Associations alors à chaque caractère C_j une variable aléatoire numérique X_j, définie de la manière suivante :

Pour une souche examinée du point de vue du caractère C_j, la variable X_j prend les valeurs :

$$\left\{ \begin{array}{l} + 1 \text{ si le caractère } C_j \text{ est présent dans la souche,} \\ - 1 \text{ si le caractère } C_j \text{ est absent de la souche,} \end{array} \right.$$

La loi de probabilité de X_j est alors (si l'on assimile les probabilités aux fréquences observées) :

$$\left\{ \begin{array}{l} \text{Prob } (X_j = + 1) = p_j/N \\ \text{Prob } (X_j = - 1) = q_j/N \end{array} \right.$$

Nous pouvons alors calculer les moments d'ordre un et deux de ces variables X_j :

moyenne de X_j :

$$\bar{x}_j = (+ 1) \times \frac{p_j}{N} + (- 1) \times \frac{q_j}{N} = \frac{1}{N} (p_j - q_j) = \frac{1}{N} (2p_j - N)$$

variance de X_j :

$$\begin{aligned} s_j^2 &= (+ 1)^2 \times \frac{p_j}{N} + (- 1)^2 \times \frac{q_j}{N} - \frac{1}{N^2} (p_j - q_j)^2 \\ &= \frac{1}{N^2} (Np_j + Nq_j - p_j^2 + 2p_jq_j - q_j^2) \\ &= \frac{1}{N^2} (p_j (N - p_j) + q_j (N - q_j) + 2p_jq_j) \\ &= \frac{1}{N^2} (p_jq_j + p_jq_j + 2p_jq_j) = \frac{4}{N^2} p_jq_j \\ s_j^2 &= \frac{4}{N^2} p_jq_j = \frac{4}{N^2} p_j (N - p_j) \end{aligned}$$

covariance de X_j et X_k : les données sur la distribution conjointe de X_j et X_k sont résumées dans le tableau II ci-dessous :

$X_k \backslash X_j$	+ 1	- 1
+ 1	$\frac{a_{jk}}{N}$	$\frac{b_{jk}}{N}$
- 1	$\frac{c_{jk}}{N}$	$\frac{d_{jk}}{N}$

- TABLEAU II -

D'où le calcul de la covariance

$$s_{jk} = (+1) \times (+1) \times \frac{a_{jk}}{N} + (-1) \times (+1) \times \frac{b_{jk}}{N} + \\ (+1) \times (-1) \times \frac{c_{jk}}{N} + (-1) \times (-1) \times \frac{d_{jk}}{N} = \bar{x}_j \bar{x}_k$$

$$s_{jk} = \frac{1}{N} (a_{jk} - b_{jk} - c_{jk} + d_{jk}) - \frac{1}{N^2} (p_j - q_j)(p_k - q_k)$$

$$s_{jk} = \frac{1}{N} (a_{jk} - b_{jk} - c_{jk} + d_{jk}) -$$

$$\frac{1}{N^2} ((a_{jk} + c_{jk}) - (b_{jk} + d_{jk})) ((a_{jk} + b_{jk}) - (c_{jk} + d_{jk}))$$

$$s_{jk} = \dots = \frac{1}{N^2} (N(a_{jk} - b_{jk}) + N(d_{jk} - c_{jk}) - (a_{jk} - b_{jk})) -$$

$$\frac{1}{N^2} ((d_{jk} - c_{jk})(d_{jk} + c_{jk}) + 2(b_{jk}c_{jk} - a_{jk}d_{jk}))$$

$$= \dots = \frac{4}{N^2} (a_{jk}d_{jk} - b_{jk}c_{jk}) \quad , \quad \text{soit en fonction de } a_{jk}, p_j, p_k :$$

$$s_{jk} = \frac{4}{N^2} (Na_{jk} - p_j p_k)$$

En résumé :

$$\bar{x}_j = \frac{1}{N} (2p_j - N) \quad ; \quad s_j^2 = \frac{4}{N^2} (Np_j - p_j^2) \quad ; \quad s_{jk} = \frac{4}{N^2} (Na_{jk} - p_j p_k)$$

La connaissance des a_{jk} et des p_i (c'est-à-dire des a_{jj}), nous permet donc d'établir la matrice des covariances correspondant aux 14 caractères observés.

Auparavant, signalons que le coefficient de corrélation de X_j et X_k , soit

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{a_{jk} d_{jk} - b_{jk} c_{jk}}{\sqrt{p_j q_j} \sqrt{p_k q_k}} \quad , \quad \text{n'est autre que le coefficient } \rho \text{ de}$$

YULE ; c'est sur la matrice de ces coefficients que DEFAYOLLE et COLOBERT ont pratiqué l'analyse factorielle.

Le vecteur moyen μ , la matrice des covariances S de l'échantillon considéré, sont alors les suivants : (avec $N = 173$)

$$t_{\mu} = \frac{1}{N} \left(\begin{array}{cccccccccccccc} 109 & -9 & -27 & 123 & -23 & -29 & -107 & 51 & 99 & 121 & -39 & 73 & 123 & 5 \end{array} \right)$$

$$S = \frac{1}{N^2} \left(\begin{array}{cccccccccccccc} 4512 & 1413 & 1990 & 1103 & 1189 & -291 & -1366 & 1335 & 546 & 552 & 587 & 476 & -108 & 1291 \\ & 7462 & 5066 & -1064 & 3684 & -3482 & -1495 & 2407 & 785 & 748 & 388 & 2889 & 493 & 3601 \\ & & 7300 & 1652 & 4213 & -3180 & -1544 & 2031 & 625 & 1033 & 299 & 3304 & 960 & 3710 \\ & & & 3700 & 1702 & 935 & -732 & 32 & 113 & 215 & 810 & 999 & 413 & 1187 \\ & & & & 7350 & -4016 & -1437 & 1461 & -512 & -645 & -8 & 2539 & 837 & 2321 \\ & & & & & 7272 & 565 & -1663 & 588 & 1007 & 1058 & -2109 & -795 & -1737 \\ & & & & & & 4620 & -1620 & -855 & -699 & -827 & -945 & -213 & -1726 \\ & & & & & & & 6832 & 1895 & 1701 & -65 & 756 & 897 & 2315 \\ & & & & & & & & 5032 & 3363 & 1960 & -293 & 113 & 1736 \\ & & & & & & & & & 3822 & 1396 & -87 & 215 & 1449 \\ & & & & & & & & & & 7102 & 409 & 810 & 1995 \\ & & & & & & & & & & & 6150 & 1518 & 3066 \\ & & & & & & & & & & & & 3700 & 1533 \\ & & & & & & & & & & & & & 7476 \end{array} \right)$$

La détermination des valeurs et vecteurs propres de la matrice T_G - détermination pratiquée avec les mêmes moyens techniques que dans les exemples précédents - a permis d'aboutir aux résultats suivants :

$\lambda_1' = 26\ 764$	$\lambda_6' = 4\ 723$	$\lambda_{11}' = 2\ 441$
$\lambda_2' = 12\ 660$	$\lambda_7' = 3\ 979$	$\lambda_{12}' = 2\ 055$
$\lambda_3' = 7\ 871$	$\lambda_8' = 3\ 275$	$\lambda_{13}' = 886$
$\lambda_4' = 6\ 141$	$\lambda_9' = 3\ 103$	$\lambda_{14}' = 487$
$\lambda_5' = 5\ 141$	$\lambda_{10}' = 2\ 804$	

$10^4\ C =$

1541	1163	-0072	-4860	-1871	1179	-2925	4875	0377	-4716	0020	-3548	0463	0504
4256	-0815	2055	2311	-2926	3731	-1878	0683	-1704	3226	1088	-1733	-1507	-5042
4418	-0511	-0761	-1543	-0612	3423	2515	1311	-1257	-0341	2356	5847	-0822	3875
0786	0988	-3938	-4610	1658	-0911	3042	-0138	0787	0267	-0992	1772	-1747	-6398
3769	-2706	-1591	-2328	-2509	-3124	-3294	-1638	0038	3294	-0565	-4507	1542	2659
-3100	3904	-2177	-2588	1574	3813	-0912	0400	-2173	5533	0691	-2072	-0587	2292
-1744	-1448	-0158	3088	0670	1511	5292	6727	0738	0719	-2538	-1338	-0031	-0084
2496	1968	5413	-1869	3264	-3975	-1051	2608	-1900	2786	-2890	1682	-0263	0350
0895	4719	2481	0860	-0820	0371	3831	-2566	0195	-2570	-0506	-2572	-5755	1159
0793	4070	2004	0171	0151	1672	3260	-1406	-1053	-1274	0841	-0418	7531	-1735
0762	4694	-4213	3116	-4713	-3532	-1287	1916	-1823	0476	-1535	1945	0392	0260
2973	-0967	-3227	1985	4310	1850	-1180	-1644	-4346	-2543	-4621	-1601	0270	0658
1123	0410	-1467	1961	3825	-3197	0493	2028	-2375	0756	7210	-2104	-0694	-0506
3765	2499	-1613	1997	2972	0843	-1718	0290	7542	1389	-0169	-0551	0674	0904

ce qui nous conduit à établir le tableau III des importances relatives des 14 facteurs pour chacune des 14 variables.

	Z ₁	Z ₂	Z ₃	Z ₄	Z ₅	Z ₆	Z ₇	Z ₈	Z ₉	Z ₁₀	Z ₁₁	Z ₁₂	Z ₁₃	Z ₁₄
X ₁	14,09	3,80	0,01	32,15	3,99	1,11	7,24	17,25	0,10	13,82	0,00	5,11	0,01	0,03
X ₂	64,97	1,13	4,45	4,40	5,90	8,31	1,62	0,20	1,21	3,91	0,39	0,23	0,27	1,66
X ₃	71,56	0,45	0,62	2,00	0,26	7,5	2,45	1,77	0,67	0,04	1,80	0,62	0,08	1,00
X ₄	4,47	3,34	32,99	35,21	3,82	1,26	0,95	0,01	0,32	0,05	0,15	1,12	1,03	5,39
X ₅	51,73	12,61	2,71	4,53	1,40	1,27	5,21	1,00	0,00	4,14	0,11	0,00	0,27	0,47
X ₆	35,37	26,53	5,13	5,66	1,7	1,44	0,30	1,10	2,61	11,50	0,16	1,21	0,01	0,35
X ₇	17,62	5,75	0,04	12,68	0,50	2,33	24,12	0,1	0,37	0,31	3,40	0,80	0,00	0,00
X ₈	24,41	7,18	33,76	3,14	0,02	10,22	0,60	1,10	1,64	3,19	2,00	0,15	0,01	0,01
X ₉	4,26	56,03	0,03	0,90	0,09	0,13	11,01	0,27	1,02	3,68	0,12	2,70	5,33	0,13
X ₁₀	4,40	54,87	8,27	0,05	0,03	3,45	11,00	1,09	0,90	1,19	0,45	0,09	13,15	0,33
X ₁₁	2,19	39,28	19,07	8,40	16,08	8,30	0,93	1,69	1,45	0,09	0,81	1,09	0,02	0,00
X ₁₂	38,46	1,92	13,33	3,93	15,53	2,03	0,90	1,44	9,53	2,95	0,48	0,86	0,01	0,03
X ₁₃	9,12	0,58	4,58	6,38	20,23	13,05	0,26	3,64	4,73	0,43	34,30	2,46	0,12	0,03
X ₁₄	50,75	10,58	2,74	3,25	6,07	0,07	1,57	0,04	23,61	0,72	0,01	0,08	0,05	0,05

- TABLEAU III -

Contrairement à ce qui se passait pour les deux premiers exemples numériques, le tableau ci-dessus ne fait pas apparaître un modèle mathématique simple. Aucun facteur n'est particulièrement important. Tous les facteurs semblent avoir un certain rôle, chacun intervenant un peu partout, d'une manière dispersée.

En fait, si nous désirons conserver les facteurs rendant compte d'une part raisonnable de la variance de chaque variable (par exemple plus de 70 %), nous sommes amenés à conserver au moins 12 ou 13 de ces 14 facteurs ; chaque variable est exprimée en fonction d' un petit nombre d'entre eux (de 4 à 7). Chacun de ces facteurs n'est à proprement parler ni un facteur commun, ni un facteur spécifique ; il s'agit plutôt d' un facteur de groupe, intervenant généralement dans plusieurs variables à la fois, mais pas dans toutes.

Pour exprimer les 14 variables étudiées, comme fonctions linéaires de facteurs incorrélés, nous avons choisi les facteurs rendant compte d'environ 75 à 80 % de la variance de chaque variable. Ce faisant, nous n'atteignons pas la précision obtenue dans les exemples précédents ; mais c'est le seul moyen de concilier les deux aspects de la question : obtenir un modèle relativement simple, et rendre compte d'une part raisonnablement importante de la variance des variables.

Au-dessous de 70 %, nous avons estimé que la part de variance due aux facteurs est trop peu importante ; au-dessus de 80 %, le modèle mathématique obtenu est beaucoup trop compliqué (voire inextricable) pour être retenu en tant que moyen de description simple de l' ensemble des 14 variables.

Nous avons donc obtenu le schéma suivant :

$$\begin{aligned}
 X_1 - 109 &= 0,2 Z_1 - 0,5 Z_4 + 0,5 Z_8 - 0,5 Z_{10} \\
 X_2 + 9 &= 0,4 Z_1 - 0,3 Z_5 + 0,4 Z_6 \\
 X_3 + 27 &= 0,4 Z_1 + 0,6 Z_{12} \\
 X_4 - 123 &= -0,4 Z_3 - 0,5 Z_4 + 0,3 Z_7 \\
 X_5 + 23 &= 0,4 Z_1 - 0,3 Z_2 - 0,3 Z_6 + 0,3 Z_7 \\
 X_6 + 29 &= -0,3 Z_1 + 0,4 Z_2 + 0,4 Z_6 + 0,6 Z_{10} \\
 X_7 + 107 &= -0,2 Z_1 + 0,3 Z_4 + 0,5 Z_7 + 0,7 Z_8 \\
 X_8 - 51 &= 0,2 Z_1 + 0,5 Z_3 + 0,3 Z_5 - 0,4 Z_6 \\
 X_9 - 99 &= 0,5 Z_2 + 0,4 Z_7 - 0,6 Z_{13} \\
 X_{10} - 121 &= 0,4 Z_2 + 0,3 Z_7 + 0,8 Z_{13} \\
 X_{11} + 39 &= 0,5 Z_2 - 0,4 Z_3 - 0,5 Z_5 \\
 X_{12} - 73 &= 0,3 Z_1 - 0,3 Z_3 + 0,4 Z_5 - 0,4 Z_9 - 0,5 Z_{11} \\
 X_{13} - 123 &= 0,1 Z_1 + 0,4 Z_5 - 0,3 Z_6 + 0,7 Z_{11} \\
 X_{14} - 5 &= 0,4 Z_1 + 0,2 Z_2 + 0,8 Z_9
 \end{aligned}$$

D - CONCLUSIONS

Le modèle ci-dessus ne permet pas de décrire d'une manière assez simple les 14 variables. En particulier, nous ne pouvons pas réduire considérablement la dimension de l'espace contenant le nuage statistique représentant l'échantillon considéré.

Rappelons que DEFAYOLLE et COLOBERT sont parvenus, par la méthode centroïde de THURSTONE, à réduire à trois cette dimension, ce qui leur a permis, dans l'espace à trois dimensions défini par les trois facteurs trouvés, de mettre en évidence le fait que le nuage statistique est divisé en trois parties distinctes (d'où l'idée que l'espèce Streptococcus faecalis est en fait formée d'au moins trois sous-groupes).

Les résultats que nous a donnés notre méthode d'analyse factorielle ne nous permettent pas de nous livrer à une telle recherche sur la structure du nuage statistique ; celui-ci, qui d'après nos résultats est nécessairement contenu dans un espace d'au moins 13 dimensions, pourrait aussi bien être étudié directement à partir des 14 variables, sans passer par l'intermédiaire des facteurs.

Cette absence de résultat valable au sujet de la structure du nuage statistique, traduit-elle un échec de notre méthode d'analyse factorielle ? Cette méthode ayant donné de bons résultats dans les deux exemples étudiés précédemment, nous ne pouvons donner de réponse catégorique à cette question. Plusieurs explications sont possibles :

- il se peut que les 14 caractères étudiés soient suffisamment peu liés entre eux pour ne pas dépendre d'un très petit nombre de facteurs ; alors il est normal de trouver beaucoup de facteurs, chacun intervenant un peu partout, d'une manière assez dispersée.

- il se peut aussi que notre méthode d'analyse factorielle, construite pour l'étude de caractères quantitatifs, ne soit plus valable pour des caractères qualitatifs. Il faudrait alors soit revenir sur la façon de "quantifier" les caractères, soit construire une méthode d'analyse factorielle spécialement conçue pour les caractères qualitatifs. Dans quelle mesure le modèle linéaire - très utilisable pour des caractères quantitatifs - est-il valable pour des caractères qualitatifs ? N'aurait-on pas intérêt dans ce cas à remplacer l'utilisation des covariances par celle d'autres indices ?

Ce sont là des questions que nous laisserons pour le moment sans réponse.

Dans cet ordre d'idées, la même méthode d'analyse, pratiquée sur des caractères qualitatifs héréditaires connus, avec linkages, pourrait, en fonction de ses résultats, nous conduire à une appréciation plus juste de la valeur de cette méthode dans le cas de caractères qualitatifs. L'étude de tels caractères qualitatifs héréditaires connus pourrait aussi servir de banc d'essai à une méthode d'analyse factorielle éventuellement mise au point pour le cas de caractères qualitatifs.

Bien que les résultats de cet essai d'analyse factorielle soient peu éloquents, du point de vue exposé ci-dessus, nous pouvons les utiliser, d'un autre point de vue, dans l'interprétation des liaisons entre caractères, sous réserve, bien entendu, de vérifications expérimentales de cette interprétation.

Ainsi, par exemple, les caractères C_9 et C_{10} (respectivement réduction et forte réduction du lait au bleu de méthylène) dépendent tous les deux des seuls facteurs Z_2 , Z_3 , Z_7 , et Z_{13} . L'examen des importances relatives de ces facteurs montrent que celles-ci sont à peu près identiques pour chacun des deux caractères, sauf celles du facteur Z_{13} , environ deux fois plus important pour C_{10} que pour C_9 . Cette constatation peut faire penser que la réduction du lait au bleu de méthylène dépend, du point de vue existentiel, de Z_2 , Z_3 , et Z_7 , et du point de vue quantitatif, de Z_{13} .

Autre constatation : le facteur Z_1 est général ; toutefois, il n'intervient pas dans les caractères relatifs à l'action de Streptococcus faecalis en présence du lait au bleu de méthylène (C_3 , C_{10} , C_{11}), ainsi que pour C_4 (mannitol).

Notons également des faits curieux tels que ceux-ci : les résultats semblent montrer une parenté étroite entre C_1 et C_7 (respectivement appartenance au groupe D de LANCEFIELD et fermentation du raffinose) : ceux-ci ont en commun quatre facteurs Z_1 , Z_4 , Z_7 , Z_8 . On pourrait ainsi noter d'autres "parentés" entre caractères, parentés reposant sur la présence de plusieurs facteurs communs. Mais de telles parentés présumées ont-elles une réalité pratique ? C'est ce que pourraient peut-être montrer des expériences destinées à les mettre particulièrement en évidence.

CONCLUSIONS GÉNÉRALES

Il semble que l'objectif que nous nous étions fixé à l'origine de ce travail a été atteint en partie. En effet :

- d'une part, nous disposons d'une méthode d'analyse factorielle solide (elle est complètement justifiée du point de vue mathématique), et objective (on peut concevoir que, à partir des données numériques brutes, toute l'analyse peut s'effectuer automatiquement - sur ordinateur par exemple - sans aucun apport humain au cours de son déroulement) ;

- d'autre part, l'analyse factorielle pratiquée en Biométrie, sur des exemples numériques précis, a donné des résultats que nous pouvons considérer comme bons dans le cas de caractères quantitatifs, mais douteux dans le cas de caractères qualitatifs.

Ce que nous n'avons pas atteint, c'est l'interprétation biologique complète des facteurs ainsi mis en évidence. Dans les conclusions qui ont suivi chacun des exemples numériques, nous n'avons fait qu'esquisser une telle interprétation. Nous en tenant strictement aux seuls résultats donnés par l'analyse factorielle, nous ne pouvions pas nous permettre de pousser plus loin l'interprétation des facteurs, faute de moyens de vérification expérimentale de telles hypothèses.

En fonction de ces qualités et défauts, quelle peut donc être la portée de l'analyse factorielle en Biométrie ? Dans tous les cas, étant donné n caractères, on peut toujours les exprimer comme fonctions liné-

aires de quelques facteurs, plus ou moins interprétables biologiquement. Cela ne doit pas signifier que l'analyse factorielle (dans son état actuel), a valeur d'une démonstration ; il ne faut pas oublier que la pratique de l'analyse factorielle en Biométrie repose sur des hypothèses, plausibles sans doute, mais hypothèses tout de même (voir Deuxième Partie). Pour cette raison, l'analyse factorielle ne peut que donner des indications, des directions de recherche ; l'étape suivant l'analyse factorielle doit consister en l'élaboration d'expériences destinées à mettre en évidence l'existence réelle des facteurs "déterminés" par le calcul.

Une analyse du genre de celle que nous avons pratiquée sur le Puceron peut, par exemple, appeler des vérifications expérimentales d'ordre génétique, portant sur la stabilité des facteurs au cours des générations successives, ou au contraire sur des variations de ces facteurs, variations suivant plus ou moins les lois de Mendel. Des essais concluants d'une sélection pratiquée sur les facteurs (et non plus sur les caractères), seraient alors un excellent résultat pour la démonstration de l'existence réelle de ces facteurs. Ceci aurait d'ailleurs un intérêt pratique certain lorsque par exemple l'on voudrait sélectionner des individus du point de vue d'un grand nombre de caractères : il suffirait d'opérer la sélection seulement sur les quelques facteurs dont dépendent tous ces caractères.

L'analyse des mensurations de la feuille de Vigne suggère des possibilités d'utilisation de cette méthode à des études de morphologie (et morphogénèse). Quels résultats obtiendrait-on si l'on pratiquait l'analyse factorielle des mêmes caractères, mais mesurés sur des espèces différentes, dont la feuille a une forme voisine (autre variété de Vitis, Platane, Erable, etc...) ? D'autre part, comment se comportent les facteurs au cours de la croissance de la feuille ? Observe-t-on entre eux les lois de dysharmonie ? Si l'on trouvait un traitement ayant pour effet un allongement (ou un raccourcissement) significatif du pétiole, sans changement du limbe moyen, certainement la nature du facteur spécifique du pétiole serait sur le point d'être identifiée.

Quant à l'analyse de caractères qualitatifs chez Streptococcus faecalis, l'absence de "bons résultats" nous enseigne au moins que nous devons nous attendre à ne pas obtenir de bons résultats dans tous les cas. C'est là une évidence, qu'il nous paraît cependant nécessaire de souligner. Il ne faut pas prêter à l'analyse factorielle des vertus qu'elle n'a pas.

Répetons qu'elle repose, en tant qu'outil biométrique, sur certaines hypothèses, et que d'autre part la méthode utilisée ici a été conçue pour des caractères quantitatifs. Or, dans le cas de Streptococcus faecalis, nous avons pratiqué une analyse de caractères qualitatifs, lesquels ont été, pour la circonstance, "quantifiés" d'une manière sinon totalement arbitraire, du moins assez artificielle. Les résultats obtenus montrent-ils que les caractères étudiés dépendent réellement d'une grande quantité de facteurs incorréllés ? Ou bien simplement est-ce un défaut inhérent à la méthode, inapplicable au cas de caractères qualitatifs ? Nous ne saurions répondre à ces questions à la suite de ce seul essai. Néanmoins, nous pouvons conclure que cette méthode d'analyse factorielle, parfaitement valable dans le cas de caractères quantitatifs, doit être manipulée avec une extrême prudence lorsqu'il s'agit de caractères qualitatifs. Nous pensons même (ce n'est là qu'une opinion) que l'analyse "linéaire" n'est pas applicable aux caractères qualitatifs, et qu'il est nécessaire de construire une méthode spéciale pour ce cas (ou bien une méthode plus générale, englobant les deux cas) (cf. DARMOIS, 1955).

Par ailleurs, il n'en reste pas moins que nous souscrivons volontiers à l'idée émise par DEFAYOLLE et COLBERT, que l'analyse factorielle (à condition de posséder la méthode appropriée ...), pourrait être un outil objectif dans des problèmes de classifications, de comparaisons entre espèces, comparaisons mal commodes si l'on considère des caractères en grand nombre, alors qu'on peut aussi bien ne considérer que les quelques facteurs dont dépendent ces caractères.

La présente conclusion serait incomplète si nous ne remarquions pas que ce travail ouvre de nombreuses portes dans des directions autres que celle de la Biométrie :

- l'inertie statistique, introduite en Première Partie, pourrait donner lieu à une étude plus théorique (Calcul des Probabilités), notamment étude des distributions des éléments propres de la matrice des covariances d'un échantillon. Une telle étude permettrait l'établissement de tests, la construction d'intervalles de confiance, utilisables notamment en analyse factorielle où ils rendraient moins arbitraire l'appréciation de l'importance des facteurs ;

- l'inertie n'est que l'une des nombreuses notions utilisées en Méca-

nique classique. Le parallèle entre le Calcul des Probabilités et la Mécanique s'étant révélé fécond au niveau de l' inertie, peut-être la statistique, s'inspirant plus profondément de la Mécanique, pourrait-elle enrichir encore ses propres méthodes ?

- l'inertie statistique n'a été appliquée ici qu' à l'analyse factorielle. Mais il est probable qu'elle peut s' appliquer à d' autres domaines de la statistique ; nous pensons par exemple à la représentation linéaire d'un ensemble de variables aléatoires ("régressions linéaires").

A P P E N D I C E

Q U E L Q U E S A S P E C T S T E C H N I Q U E S

C H A P I T R E I

COMPARAISON DES ELEMENTS PROPRES DE QUELQUES MATRICES

Dans la théorie de l'inertie statistique, et dans son application à l'analyse factorielle linéaire, c'est à la matrice T_G (tenseur d'inertie de l'échantillon) que revient le principal rôle. Nous avons vu en effet comment, du point de vue de l'analyse factorielle, on peut interpréter les valeurs propres et les directions propres de T_G .

En fait, dans la pratique, on peut remplacer T_G par des matrices plus faciles à obtenir, et dont les éléments propres sont liés à ceux de T_G par des relations très simples.

A - COMPARAISON DES ELEMENTS PROPRES DE S (MATRICE DES COVARIANCES) ET DE T_G (TENSEUR D'INERTIE)

Soit v un vecteur propre de T_G , λ la valeur propre qui lui est associée. Donc :

$$T_G v = \lambda v$$

D'autre part, T_G est défini par la relation :

$$T_G = s_E^2 I - S. \text{ On a donc :}$$

$$(s_E^2 I - S)v = \lambda v, \text{ ou } s_E^2 Iv - Sv = \lambda v, \text{ ou } Sv = (s_E^2 - \lambda)v$$

C'est dire que v est un vecteur propre de S , avec $s_E^2 - \lambda$ comme valeur propre associée.

Réciproquement, soit w un vecteur propre de S , et soit λ' la valeur propre associée.

Donc $Sw = \lambda'w$; d'où $(s_E^2 I - S)w = (s_E^2 - \lambda')w$, ce qui signifie que w est aussi vecteur propre de T_G , avec $s_E^2 - \lambda'$ comme valeur propre associée.

En résumé :

Proposition I5 : Les deux matrices carrées symétriques T_G (tenseur centré d'inertie) et S (matrice des covariances) ont les mêmes directions propres.

Proposition I6 : Si v est un vecteur propre de T_G (donc aussi de S), et si λ est la valeur propre qui lui est associée pour T_G (respectivement S), alors la valeur propre qui lui est associée pour S (respectivement T_G) est $s_E^2 - \lambda$.

B - COMPARAISON DES ELEMENTS PROPRES DE A (MATRICE CARREE QUELCONQUE) ET DE $A^* = kA$ (k nombre réel, $k \neq 0$)

Soit A^* une matrice multiple de la matrice A : $A^* = kA$ avec $k \neq 0$.

Soit v un vecteur propre de A , λ la valeur propre associée ; donc $Av = \lambda v$, et par suite $A^*v = kAv = k\lambda v$, ce qui montre que v est aussi vecteur propre de A^* , avec $k\lambda$ comme valeur propre associée.

Réciproquement, si w est un vecteur propre de A^* , associé à la valeur propre λ^* , on a $A^*w = \lambda^*w$, c'est-à-dire $kAw = \lambda^*w$, d'où $Aw = \frac{\lambda^*}{k}w$; donc w est vecteur propre de A , avec $\frac{\lambda^*}{k}$ comme valeur propre associée.

En résumé :

Proposition 17 : Les deux matrices carrées A et kA (k nombre réel non nul) ont les mêmes directions propres.

Proposition 18 : Soit v un vecteur propre de A (donc aussi de kA) et λ est la valeur propre qui lui est associée pour A , alors la valeur propre associée pour kA est $k\lambda$.

C - APPLICATION A LA MATRICE S ET A LA MATRICE T_G

Dans le présent texte, nous avons toujours désigné par S la matrice des covariances de l'échantillon E considéré. Il est entendu que nous appelons covariance de X_j et de X_k dans l'échantillon E la quantité :

$$s_{jk} = \text{cov}(X_j, X_k) = \frac{1}{N} \sum_{i=1}^N (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$$

(N est la taille de l'échantillon, c'est-à-dire le nombre d'individus composant cet échantillon).

Cette quantité s_{jk} constitue une "estimation" de la covariance vraie, c'est-à-dire de

$$\sigma_{jk} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_j - \mu_j)(x_k - \mu_k) dP(x_j, x_k)$$

(μ_j et μ_k sont les moyennes vraies de X_j et X_k ; $\bar{x}_j$ et $\bar{x}_k$ sont des "estimations" de ces moyennes).

Certains utilisateurs de la méthode statistique préfèrent employer, pour estimer la covariance σ_{jk} , la quantité :

$$\hat{s}_{jk} = \frac{1}{N-1} \sum_{i=1}^N (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = \frac{N}{N-1} s_{jk}$$

Une telle expression est appelée une estimation "sans biais" (unbiased) de σ_{jk} , par opposition à s_{jk} qui est une estimation "avec biais" (biased). Il est hors de notre propos d'examiner les raisons théori-

ques pour lesquelles il existe plusieurs sortes d'estimations des covariances. Consulter, par exemple, CRAMER (1961), chap. I3, p. 102 et chap. I4, p. 191.

Retenons simplement qu'il existe, entre les deux estimations s_{jk} et $\hat{s}_{jk}$ de la covariance σ_{jk} , la relation :

$$\hat{s}_{jk} = \frac{N}{N-1} s_{jk}$$

et notons $\hat{S}$ la matrice formée par les covariances $\hat{s}_{jk}$. Donc $\hat{S} = \frac{N}{N-1} S$

Considérons enfin la quantité

$$\sum_{i=1}^N (x_{ji} - \bar{x}_j) (x_{ki} - \bar{x}_k) ,$$

intervenant dans l'expression de la covariance (avec ou sans biais). Cette quantité peut être appelée la "covariation" de X_j et X_k dans l'échantillon E considéré. En particulier la quantité

$$\sum_{i=1}^N (x_{ji} - \bar{x}_j)^2$$

est la "variation" de X_j dans l'échantillon.

Notons S^* la matrice formée par les covariations des variables X_j et X_k dans l'échantillon ($j = 1, 2, \dots, n$; $k = 1, 2, \dots, n$). Donc :

$$S^* = N S = (N-1) \hat{S}$$

D'après les propositions I7 et I8 ci-dessus, nous pouvons énoncer:

Proposition I9 : Les trois matrices S^* (covariations), S (covariances "biased") et $\hat{S}$ (covariances "unbiased") ont les mêmes directions propres.

Proposition 20 : Soit v un vecteur propre de S^* (donc aussi de S et de $\hat{S}$). Si λ^* est la valeur propre qui lui est associée pour S , les valeurs propres associées pour S et $\hat{S}$ sont respectivement :

$$\lambda = \frac{1}{N} \lambda^* \text{ et } \hat{\lambda} = \frac{1}{N-1} \lambda^*$$

Remarquons enfin que, comme nous avons défini le tenseur centré d'inertie $T_G = s_E^2 I - S$, nous pourrions définir, dans le cadre d'une préférence pour les estimations sans biais, le tenseur sans biais correspondant, soit :

$$\hat{T}_G = \hat{s}_E^2 I - \hat{S}, \text{ avec } \hat{s}_E^2 = \hat{s}_1^2 + \hat{s}_2^2 + \dots + \hat{s}_n^2 = \frac{N}{N-1} s_E^2,$$

$$\text{soit } \hat{T}_G = \frac{N}{N-1} T_G$$

Nous avons alors :

Proposition 21 : Les deux matrices T_G (tenseur centré d'inertie) et $\hat{T}_G$ (tenseur "unbiased") ont les mêmes directions propres.

Proposition 22 : Soit v un vecteur propre de T_G (donc aussi de $\hat{T}_G$). Si λ est la valeur propre qui lui est associée pour T_G , alors la valeur propre qui lui est associée pour $\hat{T}_G$ est $\frac{N}{N-1} \lambda$.

Notons que, dans la théorie de l'inertie statistique que nous avons présentée ci-dessus, l'utilisation de $\hat{T}_G$ au lieu de T_G obligerait à certaines modifications. La "masse" totale du nuage statistique ne serait plus l'unité, mais la quantité $\frac{N}{N-1}$. En fait, l'analyse factorielle linéaire telle que nous en avons exposé le principe, peut être pratiquée soit à partir de la matrice T_G , soit à partir de la matrice $\hat{T}_G$; les différences entre les résultats obtenus sont peu importantes, nous allons en étudier la nature ci-dessous.

D - CONSEQUENCES PRATIQUES

D'après ce qui précède, les matrices S^* (covariations), S (covariances "biased"), $\hat{S}$ (covariances "unbiased"), T_G (tenseur centré d'inertie), $\hat{T}_G$ (covariances "unbiased"), ont les mêmes directions propres. Les cinq valeurs propres associées à chaque direction propre sont reliées entre elles par des relations très simples.

Dans la pratique, on peut donc effectuer la recherche des valeurs et vecteurs propres de T_G non pas obligatoirement à partir de T_G , mais à partir de l'une quelconque de ces cinq matrices, quitte à modifier, éventuellement, les valeurs propres obtenues.

Généralement, dans l'ordre chronologique des calculs, c'est la matrice S^* (covariations) qui est obtenue la première, les autres s'en déduisant par les relations indiquées plus haut. Il est donc inutile de déterminer T_G . La recherche des éléments propres peut être effectuée dès l'obtention de S^* ; la précision des calculs peut même y gagner, puisqu'on supprime des calculs intermédiaires, notamment la division par N (ou par $N - 1$) de tous les termes de la matrice S^* .

Notons bien que, quelle que soit la matrice de départ, les directions propres étant les mêmes, la matrice C de changement de base obtenue sera la même dans chacun des cinq cas. Le modèle linéaire $X = CZ$, expression des caractères X_j en fonction des facteurs Z_k , sera donc le même dans chaque cas (en particulier, les facteurs obtenus seront les mêmes, qu'ils aient été trouvés à partir de T_G ou de $\hat{T}_G$). Seules seront modifiées plus ou moins les valeurs propres, c'est-à-dire les variances de ces facteurs (on obtiendra, suivant le cas, les variances "biased" ou les variances "un-biased").

En résumé, si v est un vecteur propre commun aux cinq matrices, et si λ^* est la valeur propre qui lui est associée pour S^* , les autres valeurs propres sont données par le tableau ci-dessous :

S^*	S	$\hat{S}$	T_G	$\hat{T}_G$
λ^*	$\frac{1}{N} \lambda^*$	$\frac{1}{N-1} \lambda^*$	$\frac{1}{N} (\text{tr} S^* - \lambda^*)$	$\frac{1}{N-1} (\text{tr} S^* - \lambda^*)$

Rappelons que la notation $\text{tr} A$, où A est une matrice carrée, désigne la "trace de A ", c'est-à-dire, par définition, la somme des éléments diagonaux de A . Ainsi : $\text{tr} S = s_E^2$; $\text{tr} \hat{S} = \hat{s}_E^2$; $\text{tr} T_G = (n-1)s_E^2$; etc.....

CHAPITRE II

UNE METHODE PRATIQUE DE DETERMINATION DES ELEMENTS PROPRES

Il sera toujours préférable, dans la mesure du possible, de faire effectuer le calcul des éléments propres de la matrice des covariances par un ordinateur électronique. Cependant, en l'absence d'un tel instrument de calcul, on pourra recourir à bien cette opération au moyen d'une simple machine à calculer de bureau ; bien entendu, la précision des résultats en souffrira, et les calculs seront d'autant plus longs que le nombre des variables sera élevé.

Nous exposons ci-dessous une méthode pratique de calcul des éléments propres d'une matrice de covariances.

A - PRINCIPES DE LA METHODE

Il s'agit d'une méthode "par itération" dont les principes sont exposés d'une manière détaillée dans l'ouvrage de E. DURAND (1962, tome 2, p. 348-50, p. 389-91, p. 402-404). La méthode s'applique aux matrices symétriques à valeurs propres réelles, ce qui est le cas pour une matrice de covariances. En résumé :

- Soit $S = S_1$ la matrice $n \times n$ considérée. On choisit un vecteur initial quelconque V_1 , à n dimensions. On calcule $W_1 = S_1 V_1$. Soit h_1 la plus grande des valeurs absolues des composantes de W_1 .

On calcule $V_2 = \frac{1}{h_1} W_1$, puis $W_2 = S_1 V_2$ et $V_3 = \frac{1}{h_2} W_2$, etc.... D'une manière générale, on obtiendra, à partir de V_{k-1} :

$$W_{k-1} = S_1 V_{k-1} \quad , \quad \text{et} \quad V_k = \frac{1}{h_{k-1}} W_{k-1}$$

(h_{k-1} étant la plus grande des valeurs absolues des composantes de W_{k-1}).

On démontre que, lorsque k tend vers l'infini, la suite des vecteurs W_k tend vers le vecteur propre X_1 associé à la plus grande valeur propre de S , et la suite des nombres h_k tend vers cette valeur propre, soit λ_1 .

Dans la pratique, il n'est pas besoin que k devienne très grand ; au bout d'un certain nombre p d'itérations, on arrive au résultat $W_p = W_{p+1}$ et on en conclut que $X_1 = W_p$, et $\lambda_1 = h_p = h_{p+1}$. Dans le cas qui nous intéresse ici (analyse factorielle), on opérera en plus la "normalisation" du vecteur X_1 , en le divisant par sa longueur (euclidienne) $\sqrt{X_1 X_1}$, on obtiendra ainsi les coefficients c_{ij} qui jouent un certain rôle dans la discussion de l'analyse factorielle.

Une fois connus la plus grande valeur propre λ_1 , et le vecteur propre associé X_1 , on opère une déflation, c'est-à-dire qu'on remplace la

matrice S_1 par la matrice
$$S_2 = S_1 - \lambda_1 \frac{X_1^t X_1}{X_1^t X_1} .$$
 Une telle matrice S_2

a la particularité suivante :

Les n valeurs propres de S_1 sont identiques à celles de S_2 , sauf la plus grande (λ_1), qui dans S_2 est remplacée par zéro. Les vecteurs propres associés aux valeurs propres communes aux deux matrices leur sont également communs.

On conçoit dans ces conditions que la deuxième valeur propre de S_1 , soit λ_2 , est la plus grande valeur propre de S_2 . Il suffit donc, pour calculer λ_2 et X_2 , d'opérer sur S_2 comme on avait opéré sur S_1 pour calculer λ_1 et X_1 . Et ainsi de suite, on opère une série de déflations et d'itérations, jusqu'à obtention de toutes les valeurs propres et des vecteurs propres correspondants.

B - UN EXEMPLE NUMERIQUE

Nous appliquons la méthode décrite ci-dessus, à l'un des cas numériques traités plus haut. Nous nous sommes limités à quatre variables, qui sont les quatre segments de l'antenne du Puceron.

Les calculs ont été effectués à l'aide d'une machine de bureau "Tétractys" (Olivetti), offrant l'avantage de l'inscription des résultats sur bande de papier.

On constatera que les calculs ne sont pas extraordinairement précis (l'égalité théorique $S X_j = \lambda_j X_j$ n'est pas rigoureusement vérifiée) ; cela tient à l'obligation de perdre un certain nombre de décimales, à cha-

que itération ou à chaque déflation. Il est cependant concevable qu' une technique pourrait être étudiée, permettant, pour ce genre de calculs , d'exploiter constamment au maximum la capacité d'une telle machine à calculer (mais a-t-on vraiment besoin d'une extrême précision dans l'état actuel de l'analyse factorielle ?).

La matrice des covariances est :

$$S = S_1 = \begin{pmatrix} 1859 & 1198 & 804 & 524 \\ 1198 & 1120 & 580 & 544 \\ 804 & 580 & 509 & 381 \\ 524 & 544 & 381 & 864 \end{pmatrix}$$

détermination de λ_1 et de X_1

Nous choisissons comme vecteur initial le vecteur :

$$V_1 = \begin{pmatrix} 1,0000 \\ 0,0000 \\ 0,0000 \\ 0,0000 \end{pmatrix} ;$$

à partir de V_1 , les différentes étapes de l'itération sont consignées dans le tableau I

	1	2	3	4	5	6	7
V	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	0,0000	0,6444	0,7431	0,7555	0,7577	0,7579	0,7580
	0,0000	0,4324	0,4814	0,4882	0,4895	0,4896	0,4895
	0,0000	0,2818	0,4101	0,4371	0,4425	0,4436	0,4438
W	1859	3126	3351	3385	3392	3393	3393
	1198	2323	2532	2565	2571	2572	2572
	804	1505	1636	1657	1661	1661	1661
	524	1282	1465	1498	1505	1506	1506

On a $W_7 = W_6$; cela signifie que le vecteur W_7 (qui par construction est égal à SV_7), est proportionnel à V_7 ; le coefficient de proportionnalité est la plus grande des composantes de W_7 , soit 3 393.

Donc : la plus grande valeur propre de S est $\lambda_1 = 3\,393$; la direction propre associée est la direction du vecteur

$$W_7 = \begin{pmatrix} 3\,393 \\ 2\,572 \\ 1\,661 \\ 1\,506 \end{pmatrix} ; \text{ la longueur de ce vecteur } W_7 \text{ est } \sqrt{23\,153\,000} = 4\,812 ; \text{ le}$$

vecteur unitaire de même direction que W_7 est donc :

$$\frac{1}{4812} W_7 = X_1 = \begin{pmatrix} 0,7051 \\ 0,5345 \\ 0,3452 \\ 0,3130 \end{pmatrix}$$

détermination de λ_2 et de X_2

$$X_1^t X_1 = \begin{pmatrix} 0,4972 & 0,3769 & 0,2433 & 0,2206 \\ 0,3769 & 0,2857 & 0,1845 & 0,1672 \\ 0,2433 & 0,1845 & 0,1191 & 0,1080 \\ 0,2206 & 0,1672 & 0,1080 & 0,0979 \end{pmatrix}$$

$$\text{d'où } S_2 = S_1 - \lambda_1 X_1^t X_1 = \begin{pmatrix} 172 & -81 & -22 & -224 \\ -81 & 151 & -46 & -23 \\ -22 & -46 & 105 & 15 \\ -224 & -23 & 15 & 532 \end{pmatrix}$$

λ_2 , deuxième valeur propre de S , est aussi la plus grande valeur propre de S_2 . Nous déterminons λ_2 à partir de S_2 , de la même manière que, ci-dessus, nous avons déterminé λ_1 à partir de $S_1 = S$ (tableau II).

	1	2	3	4	5	6	7
V	1,0000	0,7679	0,5567	0,5061	0,4915	0,4875	0,4852
	0,0000	-0,3616	-0,1277	-0,0612	-0,0388	-0,0327	-0,0296
	0,0000	-0,0982	-0,0373	-0,0382	-0,0419	-0,0436	-0,0452
	0,0000	-1,0000	-1,0000	-1,0000	-1,0000	-1,0000	-1,0000
W	172	388	331	317	313	311	311
	-81	-89	-40	-25	-21	-19	-19
	-22	-26	-25	-27	-28	-29	-29
	-224	-697	-654	-645	-642	-641	-641

- TABLEAU II -

Ici encore, on a $W_7 = W_6$; donc : la valeur propre λ_2 est égale à 641 ; le vecteur

$$W_7 = \begin{pmatrix} 311 \\ -19 \\ -29 \\ -641 \end{pmatrix}$$

est vecteur propre de S_2 , donc aussi de S_1 , associé à la même valeur propre λ_2 .

La longueur de W_7 est 713,3, par conséquent nous pouvons prendre comme vecteur unitaire de même direction que W_7 le vecteur :

$$\frac{1}{713,3} W_7 = X_2 = \begin{pmatrix} 0,4360 \\ -0,0266 \\ -0,0407 \\ -0,8986 \end{pmatrix}$$

Vérifions que SX_2 est sensiblement égal à $\lambda_2 X_2$:

$$SX_2 = \begin{pmatrix} 275 \\ -20 \\ -28 \\ -578 \end{pmatrix} ; \quad \lambda_2 X_2 = \begin{pmatrix} 279 \\ -17 \\ -26 \\ -576 \end{pmatrix}$$

détermination de λ_3 et de X_3 :

$$X_2 {}^t X_2 = \begin{pmatrix} 0,1901 & -0,0116 & -0,0177 & -0,3918 \\ -0,0116 & 0,0007 & 0,0011 & 0,0239 \\ -0,0177 & 0,0011 & 0,0017 & 0,0366 \\ -0,3918 & 0,0239 & 0,0366 & 0,8075 \end{pmatrix}$$

$$\text{d'où } S_3 = S_2 - \lambda_2 X_2 {}^t X_2 = \begin{pmatrix} 50 & -74 & -11 & 27 \\ -74 & 151 & -47 & -38 \\ -11 & -47 & 104 & -8 \\ 27 & -38 & -8 & 14 \end{pmatrix}$$

En utilisant toujours la même méthode, et en partant du même vecteur initial, on trouvera :

$$W_{10} = W_3 = \begin{pmatrix} 100 \\ -212 \\ 78 \\ 51 \end{pmatrix}$$

d'où la valeur propre $\lambda_3 = 212$. Nous prendrons comme vecteur propre unitaire associé le vecteur :

$$\frac{1}{252,25} W_{10} = X_3 = \begin{pmatrix} 0,3964 \\ -0,8404 \\ 0,3092 \\ 0,2022 \end{pmatrix}$$

(la longueur de W_{10} est 252,25).

Vérifions que SX_3 est sensiblement égal à $\lambda_3 X_3$:

$$SX_3 = \begin{pmatrix} 85 \\ -177 \\ 66 \\ 43 \end{pmatrix} ; \quad \lambda_3 X_3 = \begin{pmatrix} 84 \\ -178 \\ 66 \\ 43 \end{pmatrix}$$

détermination de λ_4 et de X_4 :

$$X_3^t X_3 = \begin{pmatrix} 0,1571 & -0,3331 & 0,1226 & 0,0802 \\ -0,3331 & 0,7063 & -0,2599 & -0,1699 \\ 0,1226 & -0,2599 & 0,0956 & 0,0625 \\ 0,0802 & -0,1699 & 0,0625 & 0,0409 \end{pmatrix}$$

$$\text{d'où } S_4 = S_3 - \lambda_3 X_3^t X_3 = \begin{pmatrix} 17 & -3 & -37 & 10 \\ -3 & 1 & 8 & -2 \\ -37 & 8 & 84 & -21 \\ 10 & -2 & -21 & 5 \end{pmatrix}$$

De là on déduit, après calculs, que :

$$W_4 = W_5 = \begin{pmatrix} 47 \\ -10 \\ -106 \\ 27 \end{pmatrix} ; \text{ d'où la valeur propre } \lambda_4 = 106.$$

Nous prendrons comme vecteur propre unitaire associé le vecteur :

$$\frac{1}{119,48} W_5 = X_4 = \begin{pmatrix} 0,3934 \\ -0,0837 \\ -0,8872 \\ 0,2260 \end{pmatrix}$$

Vérifions que SX_4 est sensiblement égal à $\lambda_4 X_4$:

$$SX_4 = \begin{pmatrix} 36 \\ -19 \\ -99 \\ 19 \end{pmatrix} ; \quad \lambda_4 X_4 = \begin{pmatrix} 42 \\ -9 \\ -94 \\ 24 \end{pmatrix}$$

En conclusion, nous retiendrons les résultats suivants :

$$\lambda_1 = 3\,393 \quad ; \quad \lambda_2 = 641 \quad ; \quad \lambda_3 = 212 \quad ; \quad \lambda_4 = 106 \quad ;$$

$$C = \begin{pmatrix} 0,7051 & 0,4360 & 0,3964 & 0,3934 \\ 0,5345 & -0,0266 & -0,8404 & -0,0837 \\ 0,3452 & -0,0407 & 0,3092 & -0,8872 \\ 0,3130 & -0,8986 & 0,2022 & 0,2260 \end{pmatrix}$$

CHAPITRE III

RECAPITULATION DU DEROULEMENT DE L'ANALYSE FACTORIELLE

Les données de base sont les valeurs numériques de n caractères quantitatifs $X_1, X_2, \dots, X_n$, mesurés sur N individus composant un échantillon.

x_{ji} désigne la valeur prise par le caractère X_j sur le i -ième individu de l'échantillon.

A - MOYENNES, VARIANCES ET COVARIANCES

Calculer le vecteur moyen

$$t_{\mu} = (\bar{x}_1 \quad \bar{x}_2 \quad \dots \quad \bar{x}_n)$$

et la matrice des covariances

$$S = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1n} \\ & s_2^2 & \dots & s_{2n} \\ & & \ddots & \vdots \\ & & & s_n^2 \end{bmatrix}$$

avec

$$\bar{x}_j = \frac{1}{N} \sum_{i=1}^N x_{ji} ; s_j^2 = \frac{1}{N} \sum_{i=1}^N (x_{ji} - \bar{x}_j)^2 ; s_{jk} = \frac{1}{N} \sum_{i=1}^N (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$$

B - ELEMENTS PROPRES DE LA MATRICE DES COVARIANCES

Calculer les n valeurs propres $\lambda_1', \lambda_2', \dots, \lambda_n'$, et les n vecteurs propres unitaires correspondants $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$, de la matrice S .

Ces vecteurs (colonnes) forment la matrice C de changement de base :

$$C = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix}$$

C - IMPORTANCE DES n FACTEURS

A partir de la matrice C, des variances s_j^2 , et des valeurs propres λ_k' , calculer l'importance relative de chaque facteur Z_k pour chaque variable X_j , soit

$$\frac{a_{kj}^2 \lambda_k'}{s_j^2}$$

Il sera commode de présenter ces n nombres sous forme d'un tableau carré du type suivant :

	Z_1	Z_2		Z_n
X_1	• • •			
X_2	• • •			
•				
•				
•				
X_n	• • •			

D - DISCUSSION

Pour chaque variable X_j , classer les facteurs par ordre d'importance décroissante. Additionner ensuite les importances relatives, à partir

de la plus grande, jusqu'à obtenir un total raisonnablement voisin de 100 % (par exemple, jusqu'au facteur permettant de dépasser 95 %). Les facteurs mis en jeu dans ce total seront conservés, les autres seront éliminés ; ce principe peut toutefois subir quelques petites variations : on peut éliminer encore certains des facteurs retenus selon ce critère, ceci après une discussion tout-à-fait particulière au cas envisagé, et faisant appel à des arguments qui sortent du cadre strict de l'analyse factorielle en tant que méthode de calcul.

Contrairement aux calculs qui la précèdent, cette discussion est la partie "nuançable" de l'analyse factorielle. On ne peut établir de critères absolument rigoureux pour cette discussion, car elle dépend pour une part de la nature du problème envisagé, de son contexte, et non pas seulement des données numériques analysées.

B I B L I O G R A P H I E

Les ouvrages et articles, cités ou non dans le travail précédent, et qui nous paraissent présenter un intérêt technique ou documentaire du point de vue de l'analyse factorielle et de son utilisation en Biométrie, ont été classés en plusieurs catégories, suivant l'aspect du problème auquel ils se rapportent.

I - TECHNIQUES MATHÉMATIQUES

DURAND (E.) - 1961 - Solutions numériques des équations algébriques.

Paris, Masson Ed, tome 2 : Valeurs propres des matrices.

PERES (J.) - 1953 - Mécanique générale.

Paris, Masson Ed., chap. 2, § I : Moments d'inertie. Ellipsoïdes d'inertie.

PISOT (C.) et ZAMANSKY (M.) - 1961 - Mathématiques générales.

Paris, Dunod, Livre II : Algèbre.

II - PROBABILITES, STATISTIQUES

ANDERSON (T.W.) - 1958 - An introduction to multivariate statistical analysis.

New-York, John Wiley & sons.

CRAMER (H.) - 1951 - Mathematical methods of statistics.
Princeton, Princeton University Press.

CRAMER (H.) - 1961 - The elements of Probability Theory.
New-York, John Wiley & sons.

FISHER (R.A.) and YATES (F.) - 1957 - Statistical Tables for Use in Biological, Agricultural and Medical Research.
Edinburgh, Oliver and Boyd.

LAWLEY (D.N.) - 1956 - Tests of significance for the latent roots of covariance and correlation matrices.
Biometrika, 43, 128-136.

LOEVE (M.) - 1955 - Probability Theory. Foundations. Random Sequences.
New-York, D. Van Nostrand Co., Inc.

III - METHODES D'ANALYSE FACTORIELLE

BARTLETT (M.S.) - 1950 - Tests of significance in factor analysis.
Brit. J. Psych., Stat. sect., 3, 77-85.

BARTLETT (M.S.) - 1951 - A further note on tests of significance in factor analysis.
Brit. J. Psych., Stat. sect., 4, 1-2.

BURT (C.) - 1937 - Method of factor analysis with and without successive approximations.
Brit. J. of Ed. Psych., 7, 172.

BURT (C.) - 1949 - Alternative methods of factor analysis.
Brit. J. Psych., Stat. sect., 2, 98.

BURT (C.) - 1955 - L'Analyse Factorielle : méthodes et résultats.
Colloques du C.N.R.S., 58, 80-92 (ouvrage cité).

C.N.R.S. - 1955 - L'Analyse Factorielle et ses applications.
Colloques internationaux du C.N.R.S., 58, Paris, Ed.C.N.R.S.

DARMOIS (G.) - 1955 - Observations théoriques sur l'analyse factorielle linéaire et générale. Colloques du C.N.R.S., 58, 295-304.

- GIRSHICK (M.A.) - 1936 - Principal components.
J. Amer. Stat. Assoc., 31, 519-528.
- HOEL (P.G.) - 1937 - A significance test for component analysis.
Ann. of math. stat., 8, 149-158.
- HOLZINGER (K.J.) and HARMAN (H.H.) - 1941 - Factor analysis : a synthesis of factorial methods.
Chicago, University of Chicago Press.
- HOTELLING (H.) - 1933 - Analysis of a complex of statistical variables into principal components.
J. of Educational Psychology, 24, 417-441, 498-520.
- HOTELLING (H.) - 1936 - Simplified calculation of principal components.
Psychometrika, 1, 27-35.
- KENDALL (M.G.) and SMITH (B.B.) - 1950 - Factor analysis.
J. of the Royal Stat. soc., B, 12, 60-94.
- PEARSON (K.) - 1901 - On lines and planes of closest fit to systems of points in space.
Phil. Mag., 2, sixth series, 559-572.
- SPEARMAN (C.) - 1904 - General intelligence objectively determined and measured.
Amer. J. of Psych., 15, 201-293.
- THOMSON (G.H.) - 1939 - The factorial analysis of human ability.
London, University of London Press.
Traduction française de la 5ème édition de cet ouvrage :
Paris, P.U.F., Bibliothèque Scientifique Internationale.
- THURSTONE (L.L.) - 1931 - Multiple factor analysis.
Psychological Review, 38, 406-427.
- THURSTONE (L.L.) - 1947 - Multiple factor analysis. A development and expansion of the Vectors of Mind.
Chicago, University of Chicago Press.
- THURSTONE (L.L.) - 1955 - Problèmes actuels et méthodes nouvelles en analyse factorielle.
Colloques du C.N.R.S., 58, 31-44.

WHITTLE (P.) - 1953 - On principal components and least square methods of factor analysis.

Skandin. Aktuarietidskr., 36, p. 223-239.

IV - APPLICATIONS DE L'ANALYSE FACTORIELLE

BURT (C.) - 1940 - The factors of the mind.

London, University of London Press.

BURT (C.) and BANKS (C.H.) - 1941 - A factor analysis of body measurements for British adult males.

Ann. Eugen. Lond., 13, p. 238.

DARMOIS (G.) - 1934 - Sur la théorie des deux facteurs de Spearman.

C.R. Acad. Sc., Paris, 199, p. 231-240.

DEFAYOLLE (M.) et COLOBERT (L.) - 1962 - L'espèce Streptococcus faecalis.

II - Etude de l'homogénéité par l'analyse factorielle.

Annales de l'Institut Pasteur, 103, p. 505-522.

KELLEY (H.P.) - 1954 - A factor analysis of memory ability.

Educational Testing Service, Princeton, N.J., U.S.A.

KELLEY (T.L.) - 1935 - Essential traits of mental life.

Boston, Harvard University Press.

LEDERMAN (S.) - 1955 - Application de l'analyse factorielle à l'étude de la mortalité.

Colloques du C.N.R.S., 58, p. 263-293.

MATSAKIS (J.) - 1961 - Contribution à l'étude du développement post-embryonnaire et de l'évolution de la forme de quelques crustacés isopodes.

Thèse, Université de Toulouse.

PEARCE (S.C.) et HOLLAND (D.A.) - 1961 - Analyse des composantes, outil en recherche biométrique.

Biométrie-Praximétrie, Bruxelles, Tome 2, n° 3-4 (1961), p. 159-177.

RICARD (J.) - 1957 - Recherches quantitatives sur la croissance du Coléop-

tile des graminées.

Thèse, Faculté des Sciences de l'Université de Paris.

SCHREIDER (E.) - 1952 - Analyse factorielle de quelques caractères susceptibles de définir la structure du corps.

Biotypologie, 12, p. 26-32.

SCHREIDER (E.) - 1955 - Emploi de l'analyse factorielle dans l'étude de la variabilité biologique.

Colloques du C.N.R.S., 58, p. 253-259.

SPEARMAN (C.) - 1914 - The theory of the two factors.

Psychol. Review, 21, p. 101-115.

SPEARMAN (C.) - 1928 - The substructure of the mind.

Brit. J. Psych., 18, p. 249-261.

THURSTONE (L.L.) - 1944 - A factorial study of perception.

Psychometric Monogr., 1944, n° 4.

V - DIVERS

DE REGGI (L.) - 1963 - Recherches préliminaires sur l'hérédité de la taille chez un Puceron, Myzus Persicae (Sulz.)

Thèse, Université de Lyon.

TEISSIER (G.) - 1931 - Recherches morphologiques et physiologiques sur la croissance des Insectes.

Thèse, Université de Paris (P.U.F. édit.).

TEISSIER (G.) - 1937 - Les lois quantitatives de la croissance.

Paris, Hermann et Cie, édit.

Vu :

Les Membres du Jury,

MM. NIGON

LEGAY

BRACONNIER

Vu et approuvé

Le Doyen de la Faculté des Sciences,

J. BRACONNIER

Vu et accordé le permis d'imprimer :

Lyon, le 18 juin 1964

Le Recteur,
Président du Conseil de l' Université,

P. LOUIS

